



3 1761 10374368 8

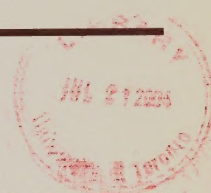
12-001



Government
Publications

131

SURVEY METHODOLOGY



Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2004

•

VOLUME 30

•


NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

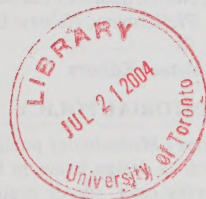
<https://archive.org/details/31761103743688>



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2004 • VOLUME 30 • NUMBER 1



Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2004

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

July 2004

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder
G.J.C. Hole
C. Patrick
R. Platek (Past Chairman)

E. Rancourt (Production Manager)
D. Roy
D. Royce
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat, Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Statistics Canada*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
J. Kovar, *Statistics Canada*
P. Lahiri, *JPSM, University of Maryland*
G. Nathan, *Hebrew University, Israel*
D. Norris, *Statistics Canada*
D. Pfeiffermann, *Hebrew University*

J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R. Valliant, *JPSM, University of Michigan*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *Iowa State University*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, Dr. M.P. Singh, singhmp@statcan.ca (Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$30 (\$15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

SURVEY METHODOLOGY
A journal Published by Statistics Canada
Volume 30, Number 1, June 2004

CONTENTS

In This Issue	1
Waksberg Invited Paper Series	
NORMAN M. BRADBURN	
Understanding the Question-Answer Process	5
Discussion Paper	
ABDELLATIF DEMNATI and J.N.K. RAO	
Linearization Variance Estimators for Survey Data	17
Comment:	
PHILLIP S. KOTT	27
BABUBHAI V. SHAH	29
CHRIS SKINNER	30
Response from the authors	32
Regular Papers	
CARY T. ISAKI, JULIE H. TSAY and WAYNE A. FULLER	
Weighting Sample Data Subject to Independent Controls	35
D. NASCIMENTO DA SILVA and JEAN D. OPSOMER	
Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism	45
J. MICHAEL BRICK, GRAHAM KALTON and JAE KWANG KIM	
Variance Estimation with Hot Deck Imputation Using a Model	57
MICHAEL A. HIDIROGLOU and ZDENEK PATAK	
Domain Estimation Using Linear Regression	67
MICHAEL SVERCHKOV and DANNY PFEFFERMANN	
Prediction of Finite Population Totals Based on the Sample Distribution	79
LEONARDO GRILLI and MONICA PRATESI	
Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs	93
GEOFF ROWE and HUAN NGUYEN	
Longitudinal Analysis of Labour Force Survey Data	105
MARC CALLENS and CHRISTOPHE CROUX	
Contact and Cooperation in the Belgian Fertility and Family Survey	115

In This Issue

This issue of *Survey Methodology* contains the fourth in the annual invited paper series in honour of Joseph Waksberg. A brief description of the series and a short biography of Joseph Waksberg were given in the June 2001 issue of the journal. I would like to thank the members of the awards selection committee for having selected Norman Bradburn as the author of this year's Waksberg invited paper.

In his paper entitled "Understanding the Question-Answer Process", Bradburn traces the history of conceptualization of the survey process over the past couple of decades, in which concepts from social and cognitive psychology and linguistics have been applied to improving our understanding of this process, and cognitive tools and approaches have been adapted for use in formulating survey instruments. He presents a conceptual model for the survey interview, and discusses various cognitive processes in survey response such as comprehension, retrieval, answer formulation and response. In his concluding summary he outlines challenges and priorities for further research in this area.

In Demnati and Rao, the authors present an approach for obtaining Taylor linearization variance estimators that is easier to apply than the usual Taylor linearization approach. The new method leads to a unique variance estimator and is applicable in many situations and estimators. The method is illustrated for calibration estimators, estimating equations and under two-phase sampling. For calibration estimators, the calibration weight is automatically captured in the variance formulae thus justifying what is commonly done in practice. Discussions of this paper are provided by Phil Kott, Babubhai Shah, and Chris Skinner.

Isaki, Tsay and Fuller propose a new method of household weighting for the 2000 U.S. Census long form, using quadratic programming to ensure that the weighted sums of household and individual characteristics match control totals derived either from the Census short form or from the Accuracy and Coverage Evaluation (A.C.E.) study. The weights are then rounded to integer values. They propose a jackknife procedure for estimation of the variance that incorporates the effects of both rounding and the random controls from A.C.E. Results of the proposed weighting procedures are compared to the 1990 weighting procedures using the 1990 Census data.

The theoretical properties of the estimator through reweighting within cells are studied in the article by da Silva and Opsomer. In contrast with numerous other studies on the subject, which involve a response model in which the population units are homogeneous within cells, it is not necessary to correctly specify the response model. It is necessary, however, to determine an auxiliary variable that is correlated with the response probability. The proposed approach can thus be seen as non-parametric. A simulation study explores the properties of the estimator being considered under various scenarios. The authors also provide some recommendations on the size and number of reweighting cells.

Brick, Kalton and Kim deal with the estimation of variance in the presence of hot-deck imputation within imputation cells for linear estimators. Särndal's decomposition (1992) and a model for the variable of interest are used to estimate variance. The originality of the proposed approach comes from the fact that, not only are the sampled and responding units conditioned, but also the units selected at the time of imputation. The article also deals with estimation for domains and a simulation study is carried out to evaluate the proposed method when certain model assumptions do not hold.

Hidiroglou and Patak study the properties of a number of small area estimators. They classify the estimators into two types, Horvitz-Thompson and Hájek, and by the detail of auxiliary information required. Conditional and unconditional properties of the estimators are investigated both analytically and in a simulation study. They conclude that the Hájek-type estimators have the best conditional properties, both in terms of bias and coverage, but these estimators do not have the additive property and their weights are domain dependent.

In their paper, Sverchkov and Pfeiffermann develop prediction of finite population totals using a model for a variable of interest conditional on the unit not being in the sample (the sample-complement distribution) and possibly some covariates. They first describe the sample distribution and the sample-complement distribution, and then develop semi-parametric estimation of the sample complement model. A resampling procedure is proposed for mean-square error estimation. The method is illustrated by examples and it is compared to alternative approaches in a simulation study.

The article by Grilli and Pratesi considers the problem of parametric estimation for ordinal and binary models at a number of levels for informational sample plans. The authors extend the pseudo maximum likelihood method to deal with this problem. This method uses the inverse of the inclusion probabilities at each degree to weight the logarithm of the likelihood function. The estimator's properties thereby obtained are tested in a simulation study. The bootstrap method is also used to obtain a variance estimator.

Rowe and Nguyen explore longitudinal analysis using data from an overlapping panel survey, specifically, the Canadian Labour Force Survey. Successive six-month longitudinal panels can be used to provide estimates relating to cohorts of people over time, provided that cohort members can be identified in each panel. They develop a likelihood function for the longitudinal data observed in each six-month window, and show how this can be used to obtain estimates of parameters of interest. They then give an illustration of this approach for estimating transition probabilities between employment states and validate it by comparing simulated and observed data.

Finally, in a paper somewhat related to Bradburn's, Callens and Croux look at individual level and municipality level predictors of contact and cooperation in the Belgian Fertility and Family Survey using multilevel logistic regression models. They discuss some social theory models for contact and cooperation that imply an important role for different indicators, and then fit models using data from the survey. Their qualitative findings, in particular with respect to socio-economic status (SES) indicators, seem to conflict with the results of similar studies in the literature. In this study, SES was found to be positively related to cooperation. Some possible explanations of the observed results are offered.

M.P. Singh

Waksberg Invited Paper Series

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic in the field of survey methodology. The author receives a cash award, made possible through a grant from Westat in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially and managed by the *American Statistical Association*. The author of the paper is selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*.

The author of the Waksberg paper is announced at the annual Joint Statistical Meeting during the American Statistical Association Presidential Address and Awards session. In this session, recipients of awards such as Section, Chapter, Continuing Education-Excellence and other co-sponsored awards are congratulated. In particular, the Waksberg Award for outstanding contributions in the theory and practice of survey methodology is highlighted. Finally, the winner of the Waksberg award appears in the Awards program booklet.

Previous Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)

Nominations:

Nominations of individuals to be considered as authors or suggestions for topics should be sent by December 3, 2004 to the chair of the committee, David Bellhouse by e-mail at: bellhouse@stats.uwo.ca or by fax (519) 661-3813.

2004 WAKSBERG INVITED PAPER

Author: Norman M. Bradburn

Norman Bradburn is the Tiffany and Margaret Blake Distinguished Service Professor Emeritus in the University of Chicago. He has spent most of his career as a survey methodologist at the National Opinion Research Center (NORC) at the University of Chicago where he is currently a Senior Fellow. His research has concentrated on the study of non-sampling errors in surveys with particular emphasis on the cognitive aspects of the survey question/answer process.

MEMBERS OF THE WASKBERG PAPER SELECTION COMMITTEE (2004-2005)

David R. Bellhouse, (Chair), *University of Western, Ontario*

Gordon Brackstone, *Statistics Canada, Ontario*

Wayne Fuller, *Iowa State University*

Sharon Lohr, *Arizona State University*

Past Chairs:

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

Understanding the Question-Answer Process

NORMAN M. BRADBURN¹

ABSTRACT

Survey statisticians have long known that the question-answer process is a source of response effects that contribute to non-random measurement error. In the past two decades there has been substantial progress toward understanding these sources of error by applying concepts from social and cognitive psychology to the study of the question-answer process. This essay reviews the development of these approaches, discusses the present state of our knowledge, and suggests some research priorities for the future.

KEY WORDS: Measurement errors; Response effects; Cognitive psychology; Questionnaire design.

1. INTRODUCTION

When I was in graduate school, I was deeply impressed by Gordon Allport's comment to the effect that the best way to find out something was to ask a direct question. Later, as I began to study and do research on methodological problems in sample surveys of human populations, I became more convinced of the wisdom on this remark. I have even formulated it into Bradburn's Law for Questionnaires: "Ask what you want to know, not something else."

The trouble with this law is that it is extremely difficult to put into practice for several reasons. First, it presumes that we know what we want to know. Often when we start out to construct a questionnaire, we are not sure what we want to know and use the questionnaire construction process in an iterative fashion to refine our ideas about what we want to know. Until we have a clear understanding of what we are trying to ask about, there is little hope that we will be able to ask meaningful questions.

Second, even if we know what we want to know, we need to understand how people answer questions. The complexities of human communication make it difficult to construct of single, standardized instrument that will enable us to ask our questions so that respondents will understand them in the way that we intend and that we will understand their answers in the way they intend. Belson (1968), who has done extensive studies on the comprehension of questions by respondents, estimates that even with the best-constructed questionnaires, less than half of the sample will understand the questions the way the researcher intended. He does not present any data on how well the researchers understand the responses.

Even if this estimate is too pessimistic, we are faced with a difficult problem of measurement error that comes from the question-answer process itself, rather than from sample

design or survey execution. The existence of this source of measurement error has been recognized since the beginning of scientific surveys, that is, since the development of sampling theory and its application to human populations. Unlike sampling theory, which rests on firm mathematical principles, the understanding of measurement error due to the question-answering process has not, until recently, been based on the theoretical understanding of human communication and cognition. This situation is beginning to change.

In the past two decades there has been substantial progress in the conceptualization of the survey interview applying concepts from social and cognitive psychology (Jabine, Straf, Tanur and Tourangeau 1984, Sudman and Bradburn 1974, Sudman, Bradburn and Schwarz 1996, Tourangeau, Rips and Rasinski 2000). In this essay I will review briefly the development of these approaches, discuss the present state of our knowledge regarding the question-answer process, and suggest some research priorities for the future.

Some History

The collaboration between cognitively oriented psychologists and survey researchers began about 25 years ago. Like many innovations it had many progenitors and seemed to spring up from several independent sources. One of the earliest, if not the earliest instance, was a seminar held in 1978 by the British Social Science Research Council and the Royal Statistical Society on problems in the collection and interpretation of recall data in social surveys. Particularly noteworthy was the participation of the Cambridge cognitive psychologist Alan Baddeley whose paper, "The Limitations of Human Memory: Implications for the Design of Retrospective Surveys," is perhaps the first paper by a psychologist interested in memory directly related to survey design (Baddeley 1979).

¹ Norman M. Bradburn, National Opinion Research Center, University of Chicago.

Two important events occurred in the United States in 1980. The first was a workshop convened by the Bureau of Social Science Research in connection with its work in the redesign of the National Crime Victimization Survey. This workshop brought together cognitive scientists and survey statisticians and methodologists to discuss what contributions cognitive scientists could make to understanding response errors in behavioral reports (Biderman 1980). One of the results of this conference was to stimulate some of the cognitive psychologists who participated to begin to study problems in survey questions in a laboratory setting. One of the earliest of such papers was "Since the eruption of Mt. St. Helens has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events," (Loftus and Marburger 1985) which demonstrated experimentally the value of using landmark events to improve the quality of dating events in survey reports.

The second event was the establishment of a panel on the measurement of subjective phenomena by the Committee on National Statistics. This panel produced two large volumes that reviewed a considerable amount of research on response effects involved in the measurement of subjective phenomena. It complemented the work that had been done by the earlier seminars on measuring behavior or more "objective" phenomena. (Turner and Martin 1982)

A big stimulus came in 1983 when the Committee on National Statistics with funding from NSF organized a 6-day seminar in St. Michaels, Maryland on Cognitive Aspects of Survey Methodology. Two papers, "Potential contributions of cognitive research to survey questionnaire design" (Bradburn and Danis 1984) and "Cognitive science and survey methods," (Tourangeau 1984) reviewed how new developments in cognitive psychology could contribute to survey methodology and how developments in survey methodology could contribute to the further development of cognitive psychology. The conference was extraordinarily fruitful and led to a whole new field of research in survey methodology both as applied to objective and subjective phenomena. The results of this conference were published in Jabine *et al.* (1984).

The final instance of independent work that may be thought of a progenitor of this field was a conference organized by Norbert Schwarz and his associates in Germany. Perhaps the most influential paper from this conference was the model proposed by Strack and Martin (1987) "Thinking, judging and communicating: A process account of context effects in attitude surveys." The results of the conference are published in Hippler, Schwarz and Sudman, *Social Information processing and survey methodology* (1987).

In the ensuing years, there has been a stream of research that has refined and elaborated the research agenda that

came from these early seminars. Some of the work sponsored by the Social Science Research Council is published in "Questions about questions: Inquiries into the cognitive bases of surveys" (Tanur 1992). Subsequent research has been updated in a series of volumes edited by Schwarz and Sudman (1992, 1994, 1996).

A Conceptual Approach to the Survey Interview

A survey interview is a structured social interaction between two people who play distinctive roles—the interviewer and the respondent. It has been described as a "conversation with a purpose" (Bingham and Moore 1934). The purpose, to put it succinctly, is to get a series of questions answered. In scientific surveys, these questions are usually embodied in a structured questionnaire designed by a third party, the researcher. It is this type of survey activity that I will consider, although the analysis could be extended to other, less structured interviews.

Like all social interactions among people from the same culture, there are implicit rules that influence the way the participants behave. Some of these are general and apply to all social interactions between social equals; some are general to the peculiar type of interaction we call the survey interview; some are general to this survey; and some are idiosyncratic and apply to only this particular interview. Thus we think of these rules as hierarchically organized from the most general, which apply to all survey interviews, to the particular rules involved in a particular interview.

At the most general level the interaction is governed by the rules for voluntary interactions between strangers. The interaction is initiated by one party, the interviewer, who must establish the nature of the encounter. The important elements that must be established are: 1) that it is non-threatening, that is the interviewer is not going to do any harm to the respondents; 2) the purpose of the encounter, and 3) what are the costs and benefits to the respondents if they agree to participate in the interview. The interaction is thus viewed as neutral, purposive, and worthwhile. As with any structured social interaction, it is governed by the norms related to such interactions.

What are the norms that are important for the interview? The first is mutual respect for individuals, particularly the privacy of the respondents. This principle has become an important issue regarding the protection of research participants because of a number of instances in bio-medical research where the voluntary nature of participation was not made clear. For high-risk research written consent to participate is now required. In the survey interview, however, the context of the request for an interview makes it easy for respondents to refuse if they do not wish to participate and written consent is superfluous. Asking for written consent may actually raise suspicion that the

interviewer has not been truthful about the purpose of the interview because written consent is not normally part of a conversation between strangers who have established that the interaction is non-threatening.

A second important norm is truthfulness. It is part of the role obligation of both parties to be truthful. For the interviewer, this means telling the respondent pertinent facts about the purpose of the interview, what is required of the respondents, *e.g.*, how much time it will take, whether they will need to consult records, whether the questions may be sensitive, *etc.* and to answer any questions the respondents might ask. If providing some information at the beginning of the interview might bias responses, such as who the sponsor of the research is, the information can be given at the end of the interview.

The purpose of the interview is to obtain the information required by the research. The interviewer's role is to get the desired information and the questionnaire is the principal instrument for accomplishing this task. A well-designed questionnaire makes the interviewer's job easier and minimizes the need for the interviewer to have to answer questions about the meaning of questions in the questionnaire. While interviewers need to be trained about the purpose of questions and their meaning, interviewers may become a source of uncontrolled variance if they have to interpret questions for many respondents. Interviewers need to be alert to cues that respondents are misunderstanding questions and to act to correct them. The need for many interventions by interviewers indicates a bad questionnaire.

If respondents accept the role and agree to participate in the interview, they have the obligation, under the norm of truthfulness, to answer the questions as accurately and completely as possible. This norm, however, may conflict with the general desire of individuals to be well thought of and to present themselves in a favorable light. In many surveys, we ask questions about potentially embarrassing, sensitive or even illegal behavior or unpopular attitudes. The interviewer and the questionnaire both play an important role in minimizing this conflict and reinforce the norm of truthfulness. The empirical evidence, however, suggests that even with the best trained interviewers and the best techniques of questionnaire design, it is rarely possible to prevent some overreporting of socially desirable behavior and attitudes or underreporting of undesirable attitudes and behavior (See Bradburn, Sudman and Associates 1979; Wentland and Smith 1993).

Survey data are collected under a strong norm of confidentiality. The norm is so strong that even if it is not made explicit, respondents expect that information from interviews that have the form of scientific surveys, such as public opinion polls or employee attitude surveys, will not be identified with them. Violations of this norm such as

occur with "sugging" (selling under the guise of a survey) or "frugging" (fund raising under the guise of a survey) threaten to erode public confidence in surveys and contribute to the increase in rates of refusal to participate. Unless the data are collected under "shield laws" or certificates of confidentiality that have the force of law, confidentiality promises, however, can be compromised by law enforcement activities.

Linguists have also noted that there are basic shared assumptions underlying conversations that facilitate the interactions. These have been systematically described by Grice and are referred to as Grician rules (Grice 1975, see also Sudman *et al.* 1996 for their application in surveys). According to Grice, conversations are based on a principle of "cooperativeness" which is embodied in four maxims. The maxim of quality enjoins speakers to be truthful and not to say things that they lack evidence for. The maxim of relation indicates that the utterances are relevant to the topic of the ongoing conversation. The maxim of quantity requires that speakers not repeat themselves and make the contributions to the conversation as informative as possible. The maxim of manner requires that the speakers be as clear as possible in their meaning. Thus, according to Grice, speakers are expected to be truthful, relevant, informative and clear.

These maxims apply equally to informal conversations and to interviews that have the form of a special type of conversation. Thus the questions asked by the interviewer are interpreted within the same framework, that is both questions and introductory material to questions are relevant to the topic, are supposed to be informative and clear. Violations of these maxims can lead to confusion on the part of respondents and produce response effects that are well documented. For example, violations of the maxim of relevance when questions are obscure (see for example, Schuman and Presser 1981) or deliberately about fictitious issues (Bishop, Oldendick and Tuchfarber 1986) lead to respondents trying to make sense of the question by looking to contextual cues about the meaning of the question. This produces what appears to be an erroneous response when viewed from the perspective of the researcher who does not understand the conversational assumptions of the respondents.

One of the most well documented order effects in surveys occurs when questions of differing levels of specificity occur together. When one question is general, *e.g.*, "Taking all things together, how happy are you these days?" and the other is specific, *e.g.*, "How happy is your marriage?", responses to the general question are affected by the order of the questions, while responses to the more specific question are not. The effect appears to be the result of the workings of the maxim of relevance. When the

general question comes first it is interpreted as intended, that is, respondents should include all aspects of their lives in making the judgment about their happiness. When the general question comes second after the specific question about marriage happiness, the maxim of relevance suggests that respondents should exclude from consideration their marriages because they have already reported on them. Thus, even though the question literally asks about "all things together", it is interpreted to mean "all things except those we have already asked about." It is only those things that have not been asked about that are still relevant.

What happens if the norms outlined above are not accepted in the interview either because the respondent rejects or redefines the role of respondent or does not observe the maxims of conversation? Of course the easiest form of rejection of the role of respondent is to refuse the interview altogether. Sometimes, however, a person sampled becomes a "reluctant respondent", that is, they are may feel pressured to participate in the study because of follow-up procedures, because they do not like to refuse a strong request from another person or for some other reason. In such cases they may care less about being a good respondent than just getting the interview finished. Thus they may take less time to think about questions, make less effort to recall information requested, or be less interested in a truthful answer than a "don't know" or even a false answer. Interviewers have told me that they often feel that the responses given by those that they have convinced to participate in an interview after many attempts at refusal conversion are less valid than those who participate more willingly. Extras efforts to obtain high completion rates may in fact produce less good data.

Respondents also may misunderstand the nature of the survey interview, simply want to convert it into a social conversation, or not be very skilled conversationalists, that is not abide by the Grician maxims and thus engage in an "inefficient" conversation. Such conversations are characterized by frequent asides or changes of topic, comments on topics of little or no relevance to the question at hand, relating personal anecdotes that may be triggered by some aspect of the question, or simple repetition of comments. In such cases the interviewer must politely but firmly teach the respondent the rules for the conversation and guide the respondent to keep focused on the questions in the interview. Skilled interviewers become experts in steering the conversation and, by selective reinforcement, shaping the respondents' behavior to follow the Grician maxims.

In summary, interviews take place in social contexts that have a structure governed by socially shared expectations and norms. These norms may differ from society to society and perhaps even within subcultures in the same society, but they have powerful effects on the way interviews are

conducted and the way questions are interpreted. Violations of the expectations or norms may lead to "effects" that may be interpreted as error from the perspective of the researcher. If these norms and expectations are understood, they can be used to avoid problems or to mitigate the effects.

Data could also be obtained from interviewers about how much the interview deviated from the model outlined above. Although little research has been done assessing the quality of interviews from this point of view, a fruitful area for future research could be to investigate the decline in validity of data as the conditions of the interview increasingly deviate from the ideal model.

Cognitive Processes in Survey Response

Answering questions in a survey involves considerable cognitive work on the part of respondents. Much of what underlies recent advances in understanding survey response processes derives from the application of models of information processing to the question-answering process. While there is still much work to be done before we have complete and detailed understanding of how the brain processes information, there is sufficient agreement about the general approach to serve as the basis for a better understanding of the response process.

The mind is conceptualized as a large information processing system composed of a series of component systems. The physical sensations of sound and sight enter the system in the sensory register. The sensory register has capacity limitations so that only a portion of the information is transferred to short-term memory. Attention plays a large role in determining what is brought into short-term memory. Attention is a function of an executive monitor that enables and controls the information processing system much the way that programs enable what computers do. The executive system controls the entire system through goals and plans that are organized into priorities for action.

The storehouse of the system is the long-term memory system that has a very large capacity. Working memory refers to the system in which active thinking takes place. The activity here draws on short-term memory and retrievals from long-term memory. Short-term memory has limited capacity but rapid access, while long-term memory has large capacity but is relatively slow in access. Long-term memory appears to have two rather distinct subsystems, semantic memory and episodic memory, although this distinction is not universally agreed upon. Semantic memory refers to memory associated with vocabulary, language structure, rules and abstract knowledge, while episodic memory refers to memory for events that took place in time and space.

Information is represented as a list of features or concepts that are linked together in networks. Information is stored in memory in structures that are hierarchically organized with more general concepts being higher in the structure than more discrete instances of the concept or distinct features. The term “schema” is sometimes used to refer to larger, more complex shared and/or overlearned structures that organize our thoughts on familiar topics and may be retrieved as a whole rather than as individual parts.

Language is the medium through which information is primarily communicated and thus information, to be available for communication, must be associated with a linguistic code. The exact relationship between language and thought and whether or not all thoughts have verbal representation are still subjects of debate. It is clear, however, that meaning is encoded somehow in language and these codes play an important role in the acquisition, storage and retrieval of information. Emotion may also be part of the code, although its role is not well understood.

Knowledge structures facilitate and constrain patterns of activation in the mind. What comes to mind, that is, into consciousness, is limited and is the result of the activation of the networks. Activation is rapid but goes along pathways determined by the ways information is encoded. Encoding puts information into particular categories and structures the pathways by which the information will be retrieved. Cues are stimuli that are related to the codes and stimulate the activation of the networks. Activation is rapid but does take time. The amount of time it takes for someone to respond to a stimulus (reaction time) is often used in research as a clue to the way information is coded.

There are number of models of the question-answering process (Cannell, Miller and Oksenberg 1981; Strack and Martin 1987; Tourangeau and Rasinski 1988; Sudman *et al.* 1996;) that, while differing in details, generally agree on a series of processes respondents go through in answering questions. These processes are: 1) comprehending the meaning of the question; 2) retrieving relevant information; 3) formulating an answer; 4) formatting and editing the answer to meet the requirements of the interviewer and respondents self-presentation. While conceptually viewed as a linear sequence, it is recognized that in reality the processes occur in the flow of a conversation and that the different processes may go on in parallel or in rapid cycling back and forth. For purposes of considering the question-answer process, it is useful to consider them as if they were separate and proceeded in an orderly sequence.

Comprehension

In order to answer a question, respondents must first understand what they are being asked. The goal for the researcher is for respondents to understand the question in

the same way that the researcher does. This goal is very difficult to reach because of the many subtleties and ambiguities of language. Indeed Belson (1981), who has studied extensively respondents’ understanding of common terms such as “weekday”, “children,” “regularly” and “proportion,” found widespread misunderstanding even in questions using such common terms.

Comprehension begins with a perceptual process of interpreting a string of sounds or written symbols as words in a language that respondents understand. The string of words is “parsed” into syntactical units that are understood, that is, the meaning that is encoded in the linguistic units is extracted by a process that is still poorly understood. Many comprehension problems occur because of ambiguities arising from words that have different meanings (lexical ambiguity) or are used in different ways (structural ambiguity). For example, the question “Where is the table?” is lexically ambiguous because the word “table” can refer to an object on which things can be placed or a set of numbers arranged in a sheet of paper. The sentence “Flying planes can be dangerous” is structurally ambiguous. The interpretation depends on whether “flying” is understood as a verb or as an adjective. Structural ambiguities can be resolved by careful wording of questions. Lexical ambiguities, on the other hand, are inherent in language and are usually resolved by the context within which the sentences appear.

Context plays an important role not only in resolving ambiguities but also aids in interpreting the meaning of words that are unfamiliar. For example, a study by Schuman and Presser (1981) found that a question about the Monetary Control Bill, an obscure piece of proposed legislation, was interpreted as referring to an anti-inflationary measure when it occurred after a series of questions about inflation, but was interpreted as referring to controls of the international transfer of money when it occurred after questions dealing with the balance of payments.

The underlying psychological mechanism for these types of context effects is priming. In order to interpret the stream of sounds or written symbols, we have to draw on our semantic memory that contains the store of linguistic information that enables us to understand the languages we know. Since this is a large store of knowledge, it takes time to retrieve information, and some things will be more easily accessible than others. Those bits of information that have been recently activated are more easily accessible and will be used first to interpret what is being said or read. Priming activates thoughts or “schemata”, that is, organized thoughts about objects or concepts, so that they are more accessible to consciousness and thus more easily come into play in interpreting the questions. In the example above, previous questions have primed either thoughts about inflation or about international flows of money, so that when the

unfamiliar concept of the Monetary Control Bill is asked about, the thoughts that have been primed come more rapidly to the fore and affect the interpretation of the words.

Different meanings may be differentially accessible to different respondents because of the frequency with which they employ them in daily life. For example, Billiet (cited in Bradburn 1992, page 317) observed that, in response to the question "How many children do you have?" some respondents offered numbers between twenty and thirty. Further inspection of the data revealed that these respondents were teachers who interpreted the question to refer to the children in their classes, the meaning that was most accessible in their memories.

Information Retrieval

Once a question has been comprehended, respondents must retrieve from memory the information necessary to answer the question. In almost all cases this means retrieving the information from long-term memory. If the question is about behavior, the relevant information is likely to be stored in episodic memory. If the question is about attitudes, the relevant information is likely to be stored in semantic memory, but may require some retrieval from episodic memory.

Remembering is a process by which the memory storehouse is searched to retrieve a particular item that is being sought. If we think of memory as a big storehouse, it is clear that it must be organized in some way in order for us to be able to retrieve things from it. Just as we must label files when we put them in file drawers, so we must attach some kind of labels to information in the memory storehouse. The labeling process, often called "encoding," refers to various aspects of the information or the experience, including emotional tone, attached to the item when we stored it in memory so that we can retrieve it. (For a more complete discussion of memory models see Tourangeau *et al.* 2000, Chapter 3).

Barsalou (1988) has proposed a theory that provides a good framework for understanding how information about personal events is stored in memory. He notes that information about activities or event types in episodic memory includes not only specific events but also extensive idiosyncratic, generic knowledge about the events, that is, having a generic mental image of some types of activity, *e.g.*, visiting a pediatrician, rather than an image of a particular event, *e.g.*, going to Dr. Jones about your daughter's rash (Brewer 1986, 1994). For activities to be stored in memory, they must be comprehended. In other words they must be understood within some meaning system, usually linguistic, that brings to bear knowledge of past activities and generic knowledge about similar event types as well as specifics of the event itself and the context

within which it occurred. This complex set of information that goes into the comprehension of the event becomes integrated into the memory of the event. The comprehension process determines how the memories are encoded.

Information, such as the wording of the question and any explanatory material available to respondents at the time they are asked to recall an event, acts as retrieval cues. Retrieval cues are any words, images, emotions, *etc.* that activate or direct the memory search process. If retrieval cues do not specify the event type, *e.g.*, pediatrician visits, then the event types must be inferred before the search can begin. This inference can come from the wording of the question or from the larger context in which the question is asked, including the preceding questions or the introductory material to the survey.

Retrieval is an active process that is facilitated by cues in the question that activate the pathways of association leading to the desired information. Because information, both in episodic and semantic memory, is encoded in many different ways, the cues in the question or in the context surrounding the question including previous questions, may facilitate or constrain the activation and produce better or less good retrieval.

Retrieval takes time. One clear empirical finding is that giving respondents more time to answer questions produces more accurate reports, particularly for behavioral questions. But time is not all there is to it. Memories for events in one's life appear to be organized in event sequences (Barsalou 1988), for example, a summer vacation or a hospitalization, which are hierarchically organized. Giving respondents cues to remind them about the sequence is more effective than trying to get them to retrieve information about a specific event. For example, in questions about alcohol consumption, giving examples of the kinds of situations in which one might drink increases consumption reports.

Examples are an important aid to recall, but they are not a panacea. Giving respondents of list of magazines that they might have read improves reports of reading; a list of organizational types helps respondents remember all the organizations they belong to. While examples may help reduce omissions, they have the effect also of being direct cues for memory and result in greater reports for the types of items on the list. If an important type of activity or event is omitted from a list, the lack of a cue for that type of activity may result in underreporting. The cuing effect of question wording can scarcely be overestimated.

When thinking about retrieval, we mostly think about forgetting or failure to retrieve relevant information. Some times, however, incorrect information may be retrieved that results in overreporting behavior. The best-known example is the phenomenon observed by Neter and Waksberg (1964)

called "telescoping", that is, recalling events that took place at a time other than the time period asked about. Telescoping occurs in response to questions about behavior in a defined time period such as: "How many times have you been to the doctor in the past 6 months?" Neter and Waksberg found in analyzing data from the Consumer Expenditure Survey that when respondents reported on purchases in different reference periods, there was a systematic overreporting of purchases that came from reporting purchases made in a previous period as if they had been purchased in the period being asked about. While the phenomenon has been observed in a number of studies, there had been no cognitive explanation for it until recently.

Memory for the time of events becomes more uncertain the further back in time the event happened, even though there is no systematic bias in the reports. Telescoping results from the conjunction of two processes—rounding and bounding. Rounding refers to the fact that respondents round their estimates for when things took place in successively larger periods the further back in time an event occurred. For example, events are remembered as having occurred in "days ago" discretely up to about 7 days ago, then they are rounded to periods such as 10 days, two weeks, 4 weeks, 3 months, and 6 months ago. Bounding refers to the aspect of the question that limits the time of reports, *e.g.*, the last 6 months. The effect of this bounding is to truncate reports of events that are remembered as having occurred longer ago than 6 months. Since the variance in the memory for the dates of events becomes larger the further back the event occurred, a larger number of events will be incorrectly remembered as falling into the period the further back the events occurred. This overreporting of events from outside the period will not be offset by an underreporting of events in the near term because events cannot be reported that have not yet happened. Since there are no offsetting events remembered as occurring outside the period at the other end of the time boundary, *i.e.*, the future, the result is a net overreport. (For a full explication of the model see Huttenlocher, Hedges and Bradburn 1990).

Formulating an Answer

Taking into account the information activated by the cues provided by the questions and the context in which they are asked and retrieved from memory, respondents must formulate an answer to the question. Some information is easily accessible. For example, if the questions are about well-rehearsed topics, such as birthdates or marital status, or about topics for which the respondents have an already well-articulated position, respondents may retrieve the answers directly. They spring, as it were, fully formed from memory and can be reported directly. This kind of information we call chronically accessible.

On the other hand, if the questions are about behavior that has not been thought about recently and is not well-remembered or about attitudes that have not been well thought out or discussed, respondents must construct answers on the spot using all the information from whatever source available to them in working memory. This construction process utilizes not only chronically available information but also, importantly, information that is temporarily accessible because it has been activated by the question itself, contextual cues, previous questions, or any other aspects of the interview situation.

There are several general cognitive processes that are pervasive strategies used to process information efficiently. Assimilation and contrast are two such fundamental processes that affect communications. In the study of perception, assimilation refers to the tendency to perceive stimuli as more alike than they actually are. Contrast refers to the tendency to perceive stimuli as more different than they actually are. Applying these principles to survey answering leads to what has been called the inclusion/exclusion model (Schwarz and Bless 1992; Sudman *et al.* 1996). Information that is included in the temporary representation that respondents form of the target of the question will result in assimilation effects because the judgment required to answer the question is based on information included in the representation used. If the information is positive, the judgment will be more positive. If the information is negative, the judgment will be more negative. The size of the effect depends on the amount and extremity of the temporarily accessible information.

Previous questions may activate thoughts that are then included in the representation of topics of later questions. The impact of a given question decreases as the number of other context questions increases. For example, answering a question about marital happiness had a pronounced effect on answers to subsequent questions about general life satisfaction when respondents' marriages were the only specific life domain asked about. When respondents were asked about their leisure time and their jobs in addition to questions about their marriages before reporting on life satisfaction, the effect was significantly reduced. (Schwarz, Strack and Mai 1991).

Information that is excluded rather than included in the temporary representation of the target will lead to a contrast effect. In this case, if the information excluded is positive, the judgment will become more negative; if the information is negative, the judgment will become more positive. Similarly the size of the effect depends on the amount and extremity of the temporarily accessible information. In effect, the excluded information is subtracted from the representation of the attitude object.

Excluded information, however, may play an additional role in formulating judgments. In addition to being excluded from the representation of the target, the information may be used in constructing a standard or scale anchor. In this case we speak of comparison-based contrast effects. The effect here is not caused so much by the subtraction of the excluded information from the evaluation of the attitude target, but by the comparison of the target with some standard or evaluated on some scale.

Which of these processes drives the emergence of a contrast effect determines whether the contrast effect is limited to the single object or generalizes across related objects. If the contrast effect is based on simple subtraction, the effect is limited to that particular target. If the contrast effect is based on a comparison, the effects are apt to appear in each judgment where that standard of comparison is relevant.

An example of a contrast effect based on using information from previous questions is provided in a study by Schwarz, Muenkel and Hippler (1990). Respondents were asked to rate a number of beverages according to how "typically German" they were. When this question was preceded by a question about the frequency with which Germans drink beer or vodka, contrast effects appear in the typicality ratings. Respondents who had estimated the consumption of beer first (a high frequency item), rated wine, milk and coffee as less typical German drinks than did respondents who had estimated the consumption of vodka first (a low frequency item), thus showing a contrast effect that extended across the three target drinks. This contrast effect, however, did not appear when the preceding question was about the caloric context of beer or vodka because the information activated by this question was not relevant to a judgment about typicality.

Formatting and Editing Responses

After respondents have formulated their responses, there remains the task of fitting these answers into the response formats that the interviewer offers. Rarely in surveys does the researcher allow respondents to answer questions in a free format. Open-ended questions have a multitude of problems not least of which is the cost and difficulty of transforming free-form answers in a format that can be treated quantitatively. Today almost all questionnaires depend on closed or pre-coded questions.

Research on response alternatives is less well developed theoretically than the study of question wording and context effects. In general, the empirically observed effects are thought to stem from two sources-memory limitations and cognitive elaboration stimulated by the response alternatives.

Memory limitations create some order effects among response alternatives. Primacy and recency are two well-known effects in the memory literature. When a series of stimuli are present visually, those that come early in the series are remembered better than those later in the series (primacy). When a series of stimuli are present in an auditory mode, those that come late in the series are remembered better (recency). Thus there is an interaction between the order in which stimuli are presented and the mode by which they are presented.

The research literature has shown that there are persistent, although in general samples fairly small, primacy and recency effects in the serial position of response alternatives depending on the mode presentation. Primacy effects appear when the response alternatives are presented visually, as in show cards in personal interviewing, and recency effects appear in telephone interviewing when the respondents have to depend entirely on auditory memory for the response alternatives. More recent research (Knaeuper 1999; Schwarz and Knaeuper 2000), however, reveals that the effect is very much a function of memory capacity and is sharply increased among older respondents whose memory is poorer and who depend more on the primacy or recency of the stimuli as supported by mode of presentation. Among older respondents, the primacy/recency effects can be quite large, on the order to 20 percentage points (Schwarz and Knaeuper 2000). Among younger respondents the effects are small.

An intriguing theory to account for some observed response order effects within a question is that of cognitive elaboration. This theory draws on early work by Krosnick and Alwin (1987) and cognitive research on persuasion (Eagly and Chaiken 1993; Petty and Cacioppo 1986). This theory hypothesizes that the order and mode in which response alternatives are presented affects respondents' opportunity to elaborate on their content. Such elaboration, in turn, activates thoughts in response to the question and provides retrieval cues in response to behavioral questions. The response alternatives provide supplementary cues that activate a range of thoughts that become temporarily accessible and may become part of the answer formulation process. In effect, the response alternatives are an essential part of the question but may be processed later in time after the question itself has been processed.

The cognitive elaboration hypothesis suggests a number of complex predictions, few of which have yet been tested. One example for which there is considerable evidence, is an interaction between serial position and mode of administration in long lists. The primacy effect evident in visually presented material gives respondents time and stimulus to think more about alternatives early in the list before giving an answer. The crowding out of early alternatives by the

reading of later alternatives and recency effects evident in lists presented in an auditory mode suggest that the later alternatives can be more deeply processed cognitively. These effects are more robust than the primacy and recency effects that appear to depend more on simple memory limitations.

Once a response alternative has been chosen in the respondents' mind, the respondent may still edit the response. As mentioned earlier, the interview is a social situation and respondents may be concerned with self-presentation. There is ample evidence that social desirability is an important aspect of the response process and responses to sensitive questions may be seriously distorted by unwillingness to admit to behavior or attitudes that would put the respondent in a bad light in the interviewer's eyes or by the desire to over claim socially desirable behavior (Bradburn, Sudman and Associates 1979; Sudman and Bradburn 1974). There are several techniques for reducing social desirability bias, although there is no technique that totally and reliably eliminates it. The general strategy is to increase social distance between respondents and interviewers. This can be done by changing the mode of administration by eliminating or reducing the presence of the interviewer. Computer Assisted Personal Interviews (CAPI) which allow respondents to directly enter responses to sensitive questions into the computer as part of a face-to-face interview enable researchers to combine the benefits of a personal interview with a self-administered questionnaire. The use of audio enhanced CAPI (Audio-CAPI) which enables respondents to listen to a recorded voice reading the questions, although somewhat more expensive, overcomes literacy and language problems that might arise when respondents have to read questions from a computer screen.

Research on mode effects generally indicates that self-administration of a questionnaire, particularly in an anonymous, group setting, minimizes, but does not entirely eliminate desirability bias. Interviews done on the telephone generally produce results that are intermediate between a face-to-face interview and a totally anonymous self-administration, although the results are not entirely consistent.

In addition to reducing the social distance between interviewer and respondent by altering the mode of administration there are techniques for increasing the real or perceived anonymity of respondents that also reduce social desirability bias. For example, respondents may put their responses in a sealed envelope and mail them back to a central office so that they know that the interviewer cannot see their responses.

Another technique is the so-called random response technique, although it is more properly a random question technique (Greenberg, Abul-El, Simmons and Horvitz 1969;

Horvitz, Shah and Simmons 1967; Warner 1965). The interviewer asks two questions, one sensitive and the other non-sensitive. Both questions have the same possible answers, "yes" and "no". Which question the respondent answers is determined by a probability mechanism, such as flipping a coin or using a plastic box containing two colored beads, e.g., red and blue beads, in differing proportions, e.g., 70% red beads and 30% blue beads. The box is designed so that when it is shaken by respondents a red or a blue bead seen only by the respondent will appear in the window of the box. If the bead is red, the sensitive question is answered; if blue, the non-sensitive question is answered. The interviewer does not know which question is answered.

By using this procedure you can estimate the behavior of a group on the sensitive questions, but not that of any single individual. Thus with this method you cannot relate individual characteristics of respondents to individual behavior. If you have a very large sample, group characteristics can be related to the estimates obtained from randomized responses. For example, you could look at all the answers of young women and compare them to all the answers of men or young versus older age groups. On the whole, however much information is lost when randomized response is used.

While, compared with other methods, randomized response greatly reduces the under reporting of undesirable behavior, it does little to reduce the overreporting of desirable behavior. It also does not entirely eliminate under-reporting of undesirable behavior (Bradburn *et al.* 1979).

CONCLUSION

In this essay, I have tried to present the outlines of a social psychological approach to the understanding of the question-answer process in the survey interview. This approach draws on theory from sociology, cognitive psychology and linguistics, to present a comprehensive framework for research on response effects. Much, however, remains uncertain or unknown.

While social role theory provides a good starting point for conceptualizing the social relations among researchers, interviewers and respondents, there is much we do not know about how these roles are played by their respective actors and how they may be changing. Contemporary concerns about privacy and confidentiality of data and protection of human participants in research are changing to an unknown degree the way respondents view surveys and social research. Technology is changing respondents' ability to protect their privacy and researchers' ability to protect confidentiality of data. Response rates have been declining and greater efforts are required to convince sampled persons to respond. Interviewing is increasingly mediated by

computer-assistance, which may change the way in which respondents and interviewers interact and the way respondents view the interview situation.

The cognitive processes involved in formulating an answer are complex and not yet fully understood. The application of our understanding of fundamental cognitive processes to the study of question formulation and order goes a long way toward improving our understanding of context effects. Cognitive science is making great strides in understanding how the brain works and how we organize and process information. New knowledge in these areas grows at a rapid pace. As we learn more, many of the conceptualizations outlined in this essay will change and either shown to be wrong or greatly elaborated.

Finally there is a great challenge to linguistics. Many of the effects we have discussed in this essay occur because of ambiguities in language. Understanding how meaning is encoded in language and how we extract that meaning from spoken and written language is a formidable challenge. Perhaps more than anything else, our ability to resolve some of the most fundamental problems in questionnaire construction depends on progress in these areas.

What are the high priority areas for research? In the short run, I would concentrate on better understanding of the biasing effects of declining respondent participation, particularly on possible distortions of responses from reluctant respondents. We must develop response effect models that not only account for missing data, whether at the item level or at the whole person level, but also for response effects introduced by reluctant respondents who give only partial answers or not well-considered answers. Multiple imputation models such as those developed by Little and Rubin (1987) and latent variable approaches such as developed by O'Muircheartaigh and Moustaki (1999) are promising. More empirical work is needed on the effects of pushing people into responding who initially are unwilling to participate in a survey.

In the longer run, further research is needed on the mechanisms by which questions and answer categories stimulate cognitive elaboration and activate thoughts that are then used in answering questions. We need to know what it is about questions that cause respondents to exclude information in making a judgment as contrasted with those that stimulate them to include information when they make judgments. Progress in this area will require a close collaboration between cognitive psychologists and survey methodologists and involve both laboratory and field survey work.

In the end, however, fundamental understanding of the question-answer process will only come when we understand how meaning is communicated between human beings. Questions have meaning that we expect respondents to comprehend. We can only go so far in improving the

process of clear communication without a much deeper understanding of the basic mechanisms of communication. We need a concerted multidisciplinary effort by linguists, psychologists, statisticians, and cognitive scientists and others to crack the meaning code much as natural scientists cracked the genetic code. It is one of the grand scientific challenges of our time.

REFERENCES

- BADDELEY, A. (1979). The limitations of human memory: Implications for the design of retrospective surveys. In *The Recall Method in Social Surveys*, (Eds. L. Moss and H. Goldstein). London: NFER Publishing Co., Ltd.
- BARSALOU, L.W. (1988). The content and organization of autobiographical memories. In *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*, (Eds. U. Neisser and E. Winograd). Cambridge, England: Cambridge University Press.
- BELSON, W.A. (1968). Respondent understand of survey questions. *Polls*. 3(1), 1-13.
- BELSON, W.A. (1981). *The Design and Understanding of Survey Questions*. Aldershot, England: Gower.
- BIDERMAN, A. (1980). *Report of a Workshop on Applying Cognitive Psychology to Recall Problems of the National Crime Survey*. Washington, D.C.: Bureau of Social Science Research.
- BINGHAM, W.V.D., and MOORE, B.V. (1934). *How to Interview* (Rev. ed.). New York: Harper Collins.
- BISHOP, G.F., OLDENDICK, R.W. and TUCHFARBER, R.J. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*. 50, 240-250.
- BRADBURN, N.M. (1992). What have we learned? In *Context Effects in Social and Psychological Research*, (Eds. N. Schwarz and S. Sudman). New York: Springer-Verlag.
- BRADBURN, N.M., and DANIS, C. (1984). Potential contributions of cognitive research to survey questionnaire design. In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau). Washington, D.C.: National Academy Press.
- BRADBURN, N.M., and SUDMAN, S. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.
- BRADBURN, N.M., SUDMAN, S. and ASSOCIATES (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.
- BREWER, W.E. (1986). What is autobiographical memory? In *Autobiographical Memory*, (Ed. D.C. Rubin). Cambridge, England: Cambridge University Press.
- BREWER, W.E. (1994). Autobiographical memory and survey research. In *Autobiographical Memory and the Validity of Retrospective Reports*, (Eds. S. Schwarz and S. Sudman). New York: Springer-Verlag.

- CANNELL, C.F., MILLER, P. and OKSENBURG, L. (1981). Research on interviewing techniques. In *Sociological Methodology 1981*, (Ed. S. Leinhardt). San Francisco: Jossey-Bass.
- EAGLY, A.H., and CHAIKEN, S. (1993). *The Psychology of Attitudes*. Orlando, FL: Harcourt, Brace, Jovanovich.
- GREENBERG, B.G., ABUL-ELA, A.L., SIMMONS W.R. and HORVITZ, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*. 64, 520-539.
- GRICE, H.P. (1975). Logic and conversation. In *Syntax and Semantics 3: Speech Acts*, (Eds. P. Cole and J.L. Morgan). New York: Academic Press. 41-58.
- HIPPLER, H.J., SCHWARZ, N. and SUDMAN, S. (Eds.) (1985). *Social Information Processing And Survey Methodology*. New York: Springer-Verlag.
- HORVITZ, D.G., SHAH, B.V. and SIMMONS, W.R. (1967). The unrelated question randomized response model. In *Proceedings of the Social Statistics Section*, American Statistical Association. 65-72.
- HUTTENLOCHER, J., HEDGES, L.V. and BRADBURN, N.M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology, Learning, Memory and Cognition*. 16, 196-213.
- JABINE, T., STRAF, M., TANUR, J. and TOURANGEAU, R. (Eds.) (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, D.C.: National Academy Press.
- KNAEUPER, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly*. 63, 347-370.
- KROSNICK, J.A., and ALWIN, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Research Quarterly* 51,201-219.
- LITTLE, R., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- LOFTUS, E.F., and MARBURGER, W. (1985). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*. 11, 114-120.
- NETER, J., and WAKSBERG, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*. 59, 18-55.
- O'MUIRCHEARTAIGH, C., and MOUSTAKI, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, A*. 162, 2, 177-194.
- PETTY, R.E., and CACIOPPO, J.T. (1986). *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.
- SCHUMAN, H., and PRESSER, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic Press.
- SCHWARZ, N., and BLESS, H. (1992). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In *The Construction of Social Judgments*, (Eds. L.L. Martin and A. Tesser). Hillsdale, N.J.: Erlbaum. 217-245.
- SCHWARZ, N., and KNAEUPER, B. (2000). Cognition, aging, and self-reports. In *Cognitive Aging: A Primer*, (Eds. D.C. Park and N. Schwarz). Philadelphia: Psychology. 233-252.
- SCHWARZ, N., MUENKEL, T. and HIPPLER, H.J. (1990). What determines a perspective? Contrast effects as a function of the dimension tapped by preceding questions. *European Journal of Social Psychology*. 20, 357-361.
- SCHWARZ, N., STRACK, F. and MAI, H.F. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*. 55, 3-23.
- SCHWARZ, N., and SUDMAN, S. (Eds.) (1992). *Context Effect in Social and Psychological Research*. New York: Springer-Verlag.
- SCHWARZ, N., and Sudman, S. (Eds.) (1994). *Autobiographical Memory and the Validity of Retrospective Reports*. New York: Springer-Verlag.
- SCHWARZ, N., and SUDMAN, S. (Eds.) (1996). *Answering Questions: Methodology of Determining Cognition and Communication Processes in Survey Research*. San Francisco: Jossey-Bass.
- STRACK, F., and MARTIN, L.L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In *Social Information Processing and Survey Methodology*, (Eds. H.J. Hippler, N. Schwarz, and S. Sudman). New York: Springer-Verlag. 123-148.
- SUDMAN, S., and BRADBURN, N.M. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.
- SUDMAN, S., BRADBURN, N.M. and SCHWARZ, N. (1996). *Thinking About Answers*. San Francisco: Jossey-Bass.
- TANUR, J.M. (Ed.) (1992). *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys*. New York: Russell Sage Foundation.
- TOURANGEAU, R. (1984). Cognitive sciences and survey methods. In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau). Washington, D.C.: National Academy Press
- TOURANGEAU, R., and RASINSKI, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*. 103, 299-314.
- TOURANGEAU, R., RIPS, L.J. and RASINSKI, K. (2000). *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.
- TURNER, C.F., and MARTIN, E. (1982). *Surveys of Subjective Phenomena*. Cambridge, MA: Harvard University Press.
- WARNER, S.L. (1965). Randomized response: A survey technique for eliminating error answer bias. *Journal of the American Statistical Association*. 60, 63-69.
- WENTLAND, E.J., and SMITH, K.W. (1993). *Survey Responses: An Evaluation of Their Validity*. San Diego: Academic Press.

Linearization Variance Estimators for Survey Data

ABDELLATIF DEMNATI and J.N.K. RAO¹

ABSTRACT

In survey sampling, Taylor linearization is often used to obtain variance estimators for calibration estimators of totals and nonlinear finite population (or census) parameters, such as ratios, regression and correlation coefficients, which can be expressed as smooth functions of totals. Taylor linearization is generally applicable to any sampling design, but it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, (ii) validity under a conditional repeated sampling framework. In this paper, a new approach to deriving Taylor linearization variance estimators is proposed. It leads directly to a variance estimator which satisfies the above considerations at least in a number of important cases. The method is applied to a variety of problems, covering estimators of a total as well as other estimators defined either explicitly or implicitly as solutions of estimating equations. In particular, estimators of logistic regression parameters with calibration weights are studied. It leads to a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators. The proposed method is extended to two-phase sampling to obtain a variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

KEY WORDS: Calibration; Design weights; Estimating equations; Raking ratio estimator; Regression estimators; Two-phase sampling.

1. INTRODUCTION

Taylor linearization is a popular method of variance estimation for complex statistics such as ratio and regression estimators and logistic regression coefficient estimators. It is generally applicable to any sampling design that permits unbiased variance estimation for linear estimators, and it is computationally simpler than a resampling method such as the jackknife. However, it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators, therefore, requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, (ii) validity under a conditional repeated sampling framework. For example, in the context of simple random sampling and the ratio estimator, $\hat{Y}_R = (\bar{y}/\bar{x})X$, of the population total Y , Royall and Cumberland (1981) showed that a commonly used linearization variance estimator, $v_L = N^2(n^{-1} - N^{-1})s_z^2$, does not track the conditional variance of \hat{Y}_R given \bar{x} , unlike the jackknife variance estimator v_J . Here \bar{y} and \bar{x} are the sample means, X is the known population total of an auxiliary variable x , s_z^2 is the sample variance of the residuals $z_k = y_k - (\bar{y}/\bar{x})x_k$ and (n, N) denote the sample and population sizes. By linearizing the jackknife variance estimator, v_J , a different linearization variance estimator, $v_{JL} = (\hat{X}/\bar{x})^2 v_L$, is obtained. This variance

estimator also tracks the conditional variance as well as the unconditional variance, where $\bar{X} = X/N$ is the mean of x . As a result, v_{JL} or v_J may be preferred over v_L . Yung and Rao (1996) considered generalized regression and ratio-adjusted post-stratified estimators under stratified multistage sampling and obtained a jackknife linearization variance estimator, v_{JL} by linearizing v_J . Valliant (1993) also obtained v_{JL} for the ratio-adjusted post-stratified estimator and conducted a simulation study to demonstrate that both v_J and v_{JL} possess good conditional properties given the estimated post-strata counts. Särndal, Swensson and Wretman (1989) showed that v_{JL} is both asymptotically design unbiased and approximately model unbiased in the sense of $E_m(v_{JL}) \approx V_m(\hat{Y}_R)$, where E_m denotes model expectation and $V_m(\hat{Y}_R)$ is the model variance of \hat{Y}_R under a "ratio model": $E_m(y_k) = \beta x_k$; $k = 1, \dots, N$ and the y_k 's are independent with model variance $V_m(y_k) = \sigma^2 x_k$, $\sigma^2 > 0$. Thus, v_{JL} is a good choice from either the design-based or the model-based perspective.

Binder (1996) presented an elegant "cookbook" approach to Taylor linearization that leads directly to v_{JL} -type linearization variance estimators. He applied the method to smooth functions of estimated totals, $g(\hat{Y}_1, \dots, \hat{Y}_m)$, generalized regression estimators and the Wilcoxon rank sum statistic. To illustrate Binder's method, consider a ratio estimator

$$\hat{Y}_R = (\hat{Y}/\hat{X})X = \hat{R}X,$$

¹ Abdellatif Demnati, Social Survey Methods Division, Statistics Canada, R.H. Coats Bldg, 15th Floor, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

where $\hat{Y} = \sum_{k=1}^N d_k(s) y_k = \hat{Y}(y)$, $\hat{X} = \sum_{k=1}^N d_k(s) x_k = \hat{X}(x)$ and the $d_k(s)$ are the design weights with $d_k(s) = 0$ if the population element k is not in the sample s , e.g., $d_k(s) = (1/\pi_k) a_k(s)$ where π_k is the probability of including the element k in the sample s , $a_k(s) = 1$ if $k \in s$, $a_k(s) = 0$ otherwise, and \sum denotes summation over the population elements. The weights are assumed to provide a design unbiased estimator \hat{Y} of Y , i.e., $E(d_k(s)) = 1$ for $k = 1, \dots, N$. Now take the total differential of \hat{Y}_R to get

$$(d\hat{Y}_R) = (d\hat{R})X = \frac{X}{\hat{X}} [(d\hat{Y}) - \hat{R}(d\hat{X})], \quad (1.1)$$

and replace all the total differentials in (1.1) by deviations of estimators from their respective population parameters, e.g., $d\hat{Y}_R$ is changed to $\hat{Y}_R - Y$. Then (1.1) yields

$$\hat{Y}_R - Y \doteq \sum d_k(s) z_k - \frac{X}{\hat{X}} (Y - \hat{R}X), \quad (1.2)$$

where

$$z_k = \frac{X}{\hat{X}} (y_k - \hat{R}x_k). \quad (1.3)$$

The term $\sum d_k(s) z_k$ in (1.2) reduces to zero, but it is retained for variance estimation. On the other hand, the last term of (1.2) is ignored for variance estimation. Thus, $\hat{Y}_R - Y$ is represented as $\sum d_k(s) z_k = \hat{Y}(z)$ for the purpose of variance estimation. Denoting an unbiased variance estimator of $\hat{Y} = \hat{Y}(y)$ as $v(y)$, Binder's variance estimator of \hat{Y}_R is given by $v(z)$. The linearization variance estimator $v(z)$, obtained from (1.3), agrees with v_{JL} for simple random sampling and stratified multistage sampling if the sample is treated as if the primary sampling units are sampled with replacement. Note that the jackknife method is not applicable generally for any sampling design.

For the estimator $\hat{\theta} = g(\hat{Y}_1, \dots, \hat{Y}_m)$ of a smooth function of totals, $\theta = g(Y_1, \dots, Y_m)$, Binder's (1996) method leads to

$$\hat{\theta} - \theta \doteq \sum d_k(s) z_k + \dots$$

with

$$z_k = \sum_{i=1}^m \left(\partial g(\mathbf{a}) / \partial a_i \Big|_{\mathbf{a}=\hat{\mathbf{y}}} \right) y_{ki}, \quad (1.4)$$

where $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_m)^T$ and $\mathbf{a} = (a_1, \dots, a_m)^T$. It follows from (1.4) that the partial derivatives, $\partial g(\mathbf{a}) / \partial a_i$, are evaluated at $\hat{\mathbf{Y}}$ to obtain z_k 's, whereas in the standard method (see e.g., Andersson and Nordberg 1994) they are evaluated at $\mathbf{Y} = (Y_1, \dots, Y_m)^T$ before getting z_k and then substituting estimates for the unknown components. For example, for the ratio estimator \hat{Y}_R the term X/\hat{X} disappears from z_k in the standard procedure because X/\hat{X} becomes 1 when \hat{X} is replaced by X .

Although Binder's (1996) approach is simple and attractive, a more rigorous and broadly applicable method is needed. In section 2, we propose an alternative approach that is theoretically justifiable and at the same time leads directly to a v_{JL} -type variance estimator for general designs. We apply the method, in section 3, to a variety of problems, covering regression calibration estimators of a total Y and other estimators defined either explicitly or implicitly as solutions of estimating equations, e.g., estimators of logistic regression parameters with design weights calibrated to known auxiliary population totals. We also obtain a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators. Section 4 extends the proposed method to two-phase sampling to obtain a variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

For the case of independent and identically (iid) random variables y_1, \dots, y_n with distribution function $F(y)$, estimation of general parameters $\theta = T(F)$ has been studied extensively in the literature (see e.g., Huber 1981). A natural estimator of $\theta = T(F)$ is $\hat{\theta} = T(\hat{F})$, where $\hat{F}(y)$ is the empirical distribution function given by $\hat{F}(y) = n^{-1} \sum_{k=1}^n I(y_k \leq y)$ with $I(y_k \leq y) = 1$ if $y_k \leq y$ and $I(y_k \leq y) = 0$ if $y_k > y$. For example, if $T(F)$ is the population mean $\int y dF(y)$, then $T(\hat{F}) = \int y d\hat{F}(y) = n^{-1} \sum_{k=1}^n y_k = \bar{y}$, the sample mean. Note that \hat{F} assigns equal mass, $1/n$ to each of the sample values y_1, \dots, y_n . If T is "sufficiently regular", then $T(\hat{F})$ may be linearized near F in terms of the influence curve (or function) of $T(\cdot)$ given by

$$IC(y, F, T) = \lim_{a \rightarrow 0} [T((1-a)F + a\delta_y) - T(F)] / a, \quad (1.5)$$

where δ_y denotes the point mass 1 at y . We have

$$\begin{aligned} \sqrt{n}[T(\hat{F}) - T(F)] &= \sqrt{n} \int IC(y, F, T) d\hat{F}(y) + \sqrt{n} R_n \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^n \tilde{z}_k + \sqrt{n} R_n \end{aligned} \quad (1.6)$$

where $\tilde{z}_k = IC(y_k, F, T)$ and $\sqrt{n} R_n$ is a remainder term. If $\sqrt{n} R_n$ is asymptotically negligible in the sense that $\sqrt{n} R_n$ converges in probability to zero as $n \rightarrow \infty$ (denoted $\sqrt{n} R_n \xrightarrow{p} 0$) then it follows from (1.6) that $\sqrt{n}[T(\hat{F}) - T(F)]$ is asymptotically normal with mean 0 and variance

$$A(F, T) = \int [IC(y, F, T)]^2 dF(y), \quad (1.7)$$

noting that the terms \tilde{z}_k in (1.6) are iid random variables. As noted by Huber (1981, page 13), $\sqrt{n} R_n$ is "often" asymptotically negligible, but the proof of this property may not be easy for general functionals $T(F)$. Serfling (1980, section 6.2) gave the following two conditions for

$\sqrt{n} R_n \rightarrow_p 0$, applicable for general random variables y_1, \dots, y_n (not necessarily iid): (i) $T(\cdot)$ is “stochastically differentiable” at F ; (ii) $\sqrt{n} \sup |\hat{F}(y) - F(y)|$ is bounded in probability, where \sup is over y . Condition (ii) is satisfied in the iid case, but it may not be easy to prove (ii) for complex sampling designs. Condition (i) means that there exists a functional $T(F; F_n - F)$ such that $T(F_n) - T(F) = n^{-1} \sum_{k=1}^n T(F; \delta_{y_k} - F) + R_n$, where R_n is of lower order in probability than $\sup |F_n(y) - F(y)|$ as the latter tends to zero. This condition may not be easy to verify for general $T(\cdot)$. Serfling (1980) suggested that in practice it is more effective to analyse R_n directly using “the method of differential inequalities”.

A natural estimator of the asymptotic variance $A(F, T)$ is

$$A(\hat{F}, T) = \frac{1}{n} \sum_{k=1}^n [\text{IC}(y_k, \hat{F}, T)]^2, \quad (1.8)$$

where $\text{IC}(y, \hat{F}, T)$ is the influence curve evaluated at $F = \hat{F}$. It follows that a linearization variance estimator of $T(\hat{F})$ is

$$v_L[T(\hat{F})] = A(\hat{F}, T)/n. \quad (1.9)$$

Practical implementation of $v_L[T(\hat{F})]$ involves the computation of $\text{IC}(y_k, \hat{F}, T)$ for each specified T . The latter can be avoided by using the jackknife method. Substituting \hat{F} for F and $-1/(n-1)$ for a in (1.5), we obtain a jackknife estimator of $\text{IC}(y_k, F, T)$ as $z_{kJ} = (n-1)[T(\hat{F}) - T(\hat{F}_{-k})]$, where $\hat{F}_{-k}(y)$ is the empirical distribution function obtained when y_k is omitted. The resulting jackknife variance estimator $T(\hat{F})$ is

$$\begin{aligned} v_J[T(\hat{F})] &= \frac{1}{n(n-1)} \sum_{k=1}^n z_{kJ}^2 \\ &= \frac{n-1}{n} \sum_{k=1}^n [T(\hat{F}_{-k}) - T(\hat{F})]^2, \end{aligned} \quad (1.10)$$

see e.g., Hampel, Ronchetti, Rousseeuw and Stahel (1986, page 95). If $\text{IC}(y, F, T)$ does not depend smoothly on F , then the jackknife variance estimator may not be consistent for the variance of $T(\hat{F})$; for example, when $T(\hat{F})$ is the sample median.

Campbell (1980) attempted to extend the above results for the iid case to general sampling designs, using the design weights $d_k(s)$. The population (or census) parameter θ is now given by $\theta = T(F_N)$, where $F_N(y)$ is the population distribution function that assigns equal mass, $1/N$, to each of the N population values y_1, \dots, y_N . An empirical distribution function is given by $\hat{F}(y) = \sum_{k \in s} \tilde{d}_k(s) I(y_k \leq y)$, where $\tilde{d}_k(s) = d_k(s) / \sum_{l \in s} d_l(s)$ are the normalized design weights. Note that $\hat{F}(y)$ assigns the mass $\tilde{d}_k(s)$ to the element $k \in s$. An estimator of $\theta = T(F_N)$ is given by $\hat{\theta} = T(\hat{F})$. For example, if $T(F_N)$ is the population mean

$\int y dF_N(y)$, then $T(\hat{F}) = \int y d\hat{F}(y) = \sum_{k \in s} d_k(s) y_k / \sum_{k \in s} d_k(s)$, the design-weighted sample mean. Campbell (1980) followed the linearization (1.6) for the iid case and concluded that $\sqrt{n} [T(\hat{F}) - T(F_N)]$ is asymptotically normal with mean 0 and variance

$$\begin{aligned} A(F_N, T) &= n \text{Var} \left[\sum_{k \in s} d_k(s) \tilde{z}_k / \sum_{k \in s} d_k(s) \right] \\ &\approx n \text{Var} \left[\sum_{k \in s} d_k(s) \{(\tilde{z}_k - R)/N\} \right], \end{aligned} \quad (1.11)$$

using the approximate variance of a ratio, where $R = \sum_{k \in s} \tilde{z}_k / N$ is the population mean of \tilde{z}_k 's and $\tilde{z}_k = \text{IC}(y_k, F_N, T)$. Denoting the unbiased variance estimator of $\hat{Y} = \hat{Y}(y) = \sum_{k \in s} d_k(s) y_k$ as $v(y)$, it follows from (1.11) that a linearization variance estimator of $T(\hat{F})$ is given by

$$v_L[T(\hat{F})] = v[(z - \hat{R})/\hat{N}], \quad (1.12)$$

where

$$z_k = \text{IC}(y_k, \hat{F}, T), \quad (1.13)$$

and

$$\hat{R} = \sum_{k \in s} d_k(s) z_k / \sum_{k \in s} d_k(s). \quad (1.14)$$

To avoid the computation of z_k 's, Campbell (1980) proposed a jackknife estimator of \tilde{z}_k for each $k \in s$. It is given by

$$z_{kJ} = \frac{1 - \tilde{d}_k(s)}{\tilde{d}_k(s)} [T(\hat{F}) - T(\hat{F}_{-k})], \quad (1.15)$$

where

$$d\hat{F}_{-k}(y) = \begin{cases} \frac{d\hat{F}(y) - \tilde{d}_k(s)}{1 - \tilde{d}_k(s)} & \text{if } y = y_k \\ \frac{d\hat{F}(y)}{1 - \tilde{d}_k(s)} & \text{if } y \neq y_k. \end{cases} \quad (1.16)$$

The resulting linearization variance estimator is given by $v[(z_J - \hat{R}_J)/\hat{N}]$. Note that the proposed jackknife method is different from the customary jackknife for survey sampling. For example, for stratified multistage sampling, the customary jackknife deletes sample clusters in turn whereas the Campbell method deletes elements in turn. Also, the customary jackknife is not always applicable (e.g., unequal probability sampling without replacement) unlike the Campbell method which uses the unbiased variance estimator $v(y)$ of the total \hat{Y} for the given design and then replaces y by $(z_J - \hat{R}_J)/\hat{N}$. However, the computations involved in the Campbell method can be very heavy because it requires the computation of $T(\hat{F}_{-k})$ for each element $k \in s$; in large-scale surveys the number of sample

elements can be very large, as in the Canadian Labour Force Survey.

Deville (1999) and Berger (2002) obtained results very similar to those of Campbell (1980). Instead of using the natural probability measure \hat{F} , they considered functionals of the form $T(\hat{M})$, where \hat{M} denotes a measure that allocates the design weight $d_k(s)$ to any point y_k for k in s and zero to units k not in s . For example, $T(\hat{M}) = \int x d\hat{M}(x) = \sum d_k(s) y_k$ if the population parameter is the total $T(M) = \int x dM(x) = Y$, where the measure M allocates a unit mass to each of the N points y_k in the finite population U . Suppose that $T(\cdot)$ is of degree α in the sense that $N^{-\alpha} T(\cdot)$ tends to a limit for some $\alpha \geq 0$. Typically, $\alpha = 0$ or 1 ; for example, $\alpha = 1$ if $T(M)$ is the total Y and $\alpha = 0$ if $T(M)$ is the ratio $R = Y/X$. Deville (1999) used the following asymptotic approximation:

$$\sqrt{n} N^{-\alpha} [T(\hat{M}) - T(M)] \approx \frac{\sqrt{n}}{N} \sum (d_k(s) - 1) \tilde{z}_k, \quad (1.17)$$

where $d_k(s) = 0$ if k is not in the sample s . Further $\tilde{z}_k = \text{IT}(M; y_k)$ with IT denoting the influence function of $T(M)$ defined by

$$\text{IT}(M; y) = \lim_{t \rightarrow 0} \frac{1}{t} [T(M + t\delta_y) - T(M)]. \quad (1.18)$$

As noted earlier, it is not easy to justify the approximation (1.17) for general functionals $T(\cdot)$. Deville (1999) developed rules for evaluating $\text{IT}(M; y)$ for selected functionals $T(\hat{M})$. Berger (2002) used the jackknife method to estimate $\tilde{z}_k = \text{IT}(M, y_k)$, similar to Campbell (1980).

Noting that $\sum d_k(s) \tilde{z}_k = \hat{Y}(\tilde{z})$ it follows from (1.17) that a linearization variance estimator of $N^{-\alpha} T(\hat{M})$ is given by $N^{-2} v(\tilde{z})$. But \tilde{z}_k depends on unknown parameters and the corresponding estimator, z_k , may not be unique. For example, suppose $T(\hat{M}) = \hat{Y}_R = (\hat{Y}/\hat{X})X$, then $\alpha = 1$ and $\tilde{z}_k = y_k - R x_k$, where $R = Y/X$. In this case, two possible candidates for z_k are $z_k = y_k - \hat{R} x_k$ and $z_k = (X/\hat{X})(y_k - \hat{R} x_k)$. Thus, the choice of z_k in the presence of auxiliary information, such as a known total X , is not unique under Deville's approach. Unlike Deville's approach, our method leads to a unique choice z_k and it avoids the calculation of \tilde{z}_k to determine z_k . Our z_k satisfies desirable properties mentioned in section 1, at least in a number of important cases.

2. THE METHOD

To motivate the method, we start with a simple general case where the estimator $\hat{\theta}$ of a parameter θ can be expressed as a smooth function $g(\hat{Y})$ of estimated totals $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_i, \dots, \hat{Y}_m)^T$, where $\hat{Y}_i = \sum_{k \in U} d_k(s) y_{ik}$,

$i = 1, \dots, m$, is an estimator of the total $Y_i = \sum_{k \in U} y_{ik}$, and $\theta = g(\mathbf{Y})$ with $\mathbf{Y} = (Y_1, \dots, Y_i, \dots, Y_m)^T$. We may write $\hat{\theta}$ as $\hat{\theta} = f(\mathbf{d}(s), \mathbf{A}_y)$ and $\theta = f(\mathbf{1}, \mathbf{A}_y)$, where \mathbf{A}_y is an $m \times N$ matrix with k^{th} column $\mathbf{y}_k = (y_{k1}, \dots, y_{km})^T$, $k = 1, \dots, N$, $\mathbf{d}(s) = (d_1(s), \dots, d_N(s))^T$ and $\mathbf{1}$ is the N -vector of 1's. For example, if $\hat{\theta}$ denotes the ratio estimator $\hat{Y}_R = [(\sum d_k(s) y_k) / (\sum d_k(s) x_k)] X$, then $m = 2$, $y_{1k} = y_k$, $y_{2k} = x_k$ and $f(\mathbf{1}, \mathbf{A}_y)$ reduces to the total Y , noting that $(Y/X)X = Y$. Note that \hat{Y}_R is a function of $\mathbf{d}(s)$, \mathbf{y} and \mathbf{x} and the known total X , but we dropped X for simplicity and write $\hat{Y}_R = f(\mathbf{d}(s), \mathbf{y}, \mathbf{x})$.

Taylor linearization of $\hat{\theta}$ around \mathbf{Y} gives the approximation

$$\sqrt{n} N^{-\alpha} (\hat{\theta} - \theta) \approx \frac{\sqrt{n}}{N} \left(\partial g(\mathbf{a}) / \partial \mathbf{a} \right)^T \Big|_{\mathbf{a}=\mathbf{Y}} (\hat{\mathbf{Y}} - \mathbf{Y}) \quad (2.1)$$

where $\partial g(\mathbf{a}) / \partial \mathbf{a} = (\partial g(\mathbf{a}) / \partial a_1, \dots, \partial g(\mathbf{a}) / \partial a_m)^T$ and $N^{-\alpha} g(\cdot)$ tends to a limit for some $\alpha \geq 0$. Asymptotic normality of $\sqrt{n} N^{-\alpha} (\hat{\theta} - \theta)$ follows from (2.1), provided a central limit theorem for $\sqrt{n} N^{-1} (\hat{\mathbf{Y}} - \mathbf{Y})$ holds and $g(\cdot)$ has continuous first derivatives in a neighbourhood of the mean $\bar{\mathbf{Y}}$. Krewski and Rao (1981) justified (2.1) for stratified sampling.

Let $\check{\mathbf{Y}} = \sum b_k \mathbf{y}_k$ for arbitrary real numbers $\mathbf{b} = (b_1, \dots, b_N)^T$, and $g(\check{\mathbf{Y}}) = f(\mathbf{b}, \mathbf{A}_y) = f(\mathbf{b})$. Noting that $\hat{\mathbf{Y}} = \mathbf{A}_y \mathbf{d}(s)$ and $\mathbf{Y} = \mathbf{A}_y \mathbf{1}$, we can express (2.1) as

$$\begin{aligned} \sqrt{n} N^{-\alpha} (\hat{\theta} - \theta) &\approx \frac{\sqrt{n}}{N} \left(\partial g(\check{\mathbf{Y}}) / \partial \check{\mathbf{Y}} \right)^T \Big|_{\check{\mathbf{Y}}=\mathbf{Y}} \mathbf{A}_y (\mathbf{d}(s) - \mathbf{1}) \\ &= \frac{\sqrt{n}}{N} \sum_{k=1}^N \left(\partial f(\mathbf{b}) / \partial \check{\mathbf{Y}} \right)^T \Big|_{\mathbf{b}=\mathbf{1}} \mathbf{y}_k (d_k(s) - 1), \end{aligned} \quad (2.2)$$

noting that $\check{\mathbf{Y}} = \mathbf{Y}$ is equivalent to $\mathbf{b} = \mathbf{1}$. Now we substitute $\mathbf{y}_k = \partial \check{\mathbf{Y}} / \partial b_k \Big|_{\mathbf{b}=\mathbf{1}}$ in (2.2) to get

$$\begin{aligned} \sqrt{n} N^{-\alpha} (\hat{\theta} - \theta) &\approx \frac{\sqrt{n}}{N} \sum_{k=1}^N \left(\partial f(\mathbf{b}) / \partial b_k \right) \Big|_{\mathbf{b}=\mathbf{1}} (d_k(s) - 1) \\ &= \frac{\sqrt{n}}{N} \tilde{\mathbf{z}}^T (\mathbf{d}(s) - \mathbf{1}), \end{aligned} \quad (2.3)$$

where $\tilde{\mathbf{z}} = (\tilde{z}_1, \dots, \tilde{z}_N)^T$ with $\tilde{z}_k = \partial f(\mathbf{b}) / \partial b_k \Big|_{\mathbf{b}=\mathbf{1}}$.

A variance estimator of the right hand side of (2.3) is given by $(n/N^2) v(\tilde{\mathbf{z}})$, where $v(\tilde{\mathbf{z}})$ is the variance estimator of the estimated total $\sum d_k(s) \tilde{z}_k = \hat{Y}(\tilde{\mathbf{z}})$. Since \tilde{z}_k 's are unknown, we replace \tilde{z}_k by $z_k = \partial f(\mathbf{b}) / \partial b_k \Big|_{\mathbf{b}=\mathbf{d}(s)}$, to get $(n/N^2) v(\mathbf{z})$. Thus, a linearization variance estimator of $\hat{\theta}$ is given by

$$v_L(\hat{\theta}) = (N^{2\alpha}/N^2) v(\mathbf{z}), \quad (2.4)$$

which reduces to $v(\mathbf{z})$ if $\alpha = 1$. Note that $v_L(\hat{\theta})$ given by (2.4) is simply obtained from the formula $v(\mathbf{y})$ for $\hat{\mathbf{Y}}$ by replacing y_k by z_k for $k \in s$. Note that we do not first

evaluate the partial derivatives $\partial f(\mathbf{b})/\partial b_k$ at $\mathbf{b} = \mathbf{1}$ to get \tilde{z} and then substitute estimates for the unknown components of \tilde{z} . Our method, therefore, is similar in spirit to Binder's approach. The variance estimator $v_L(\hat{\theta})$ is valid because z_k is a consistent estimator of \tilde{z}_k .

Example 2.1 Suppose $\hat{\theta}$ is the ratio estimator $\hat{Y}_R = X[(\sum d_k(s)y_k)/(\sum d_k(s)x_k)]$ of the total Y . Then $f(\mathbf{b}) = X[(\sum b_k y_k)/(\sum b_k x_k)]$ and

$$\partial f(\mathbf{b})/\partial b_k = X \frac{y_k \sum b_k x_k - x_k \sum b_k y_k}{(\sum b_k x_k)^2}.$$

Therefore,

$$z_k = \partial f(\mathbf{b})/\partial b_k|_{\mathbf{b}=\mathbf{d}(s)} = \frac{X}{\hat{X}}(y_k - \hat{R}x_k)$$

which agrees with (1.3). Thus, our variance estimator $v_L(\hat{Y}_R)$ is identical to Binder's (1996) variance estimator, $v(z)$, noting that $\alpha = 1$.

Our derivation is simple and natural. On the other hand, in the standard linearization method, $\hat{\theta}$ is first expressed in terms of elementary components $\hat{Y}_1, \dots, \hat{Y}_m$ as $g(\hat{\mathbf{Y}})$ and the partial derivatives $\partial g(\mathbf{a})/\partial a_j$ are then evaluated at $\mathbf{a} = \mathbf{Y}$. It is interesting to note that all the components of $\hat{\mathbf{Y}}$ use the same weights $d_k(s)$ and our approach always takes first derivatives of $f(\mathbf{b})$ with respect to b_k at $\mathbf{b} = \mathbf{d}(s)$. It is not necessary to first express $\hat{\theta}$ in terms of elementary components.

3. CALIBRATION ESTIMATORS

The ratio estimator can be viewed as a calibration estimator, $\hat{Y}_R = \sum w_k(s)y_k$, with explicit weights $w_k(s) = (X/\hat{X})d_k(s)$ and satisfying the calibration constraint $\sum w_k(s)x_k = X$. Calibration estimators of a total Y of the form $\hat{Y}_w = \sum w_k(s)y_k$ with explicit weights $w_k(s)$ and satisfying the calibration constraints $\sum w_k(s)\mathbf{x}_k = \mathbf{X}$ are widely used, where $\mathbf{x}_k = (x_{1k}, \dots, x_{qk})^T$ and $\mathbf{X} = (X_1, \dots, X_q)^T$ is the vector of known totals of auxiliary variables x_{jk} , $j = 1, \dots, q$. In subsection 3.1 we consider the generalized regression (GREG) estimator and then study a general class of regression calibration estimators in subsection 3.2. Extension to estimators, $\hat{\theta}$, obtained as solutions of estimating equations is presented in subsection 3.3. The case of general calibration estimators is investigated in subsection 3.4.

3.1 Generalized Regression Estimator

The GREG estimator of total Y is given by \hat{Y}_w with calibration weights $w_k(s) = d_k(s)g_k(\mathbf{d}(s))$, where

$$g_k(\mathbf{d}(s)) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum d_k(s) c_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} c_k \mathbf{x}_k \quad (3.1)$$

with specified constants c_k and $\hat{\mathbf{X}} = \sum d_k(s)\mathbf{x}_k$ (cf., Särndal *et al.* 1989). The ratio estimator, \hat{Y}_R , is a special case with $q=1$ (i.e., scalar x_k) and $c_k = x_k^{-1}$, and $g_k(\mathbf{d}(s))$, given by (3.1), reduces to X/\hat{X} .

The GREG estimator may be expressed as a differentiable function of estimated totals. Hence, the general theory of section 2 is applicable and it remains to evaluate $z_k = \partial f(\mathbf{b})/\partial b_k|_{\mathbf{b}=\mathbf{d}(s)}$, where $f(\mathbf{b}) = \sum (b_k g_k(\mathbf{b}))y_k$ is obtained by replacing $\mathbf{d}(s)$ by \mathbf{b} in the formula for \hat{Y}_w . Noting that $\partial \mathbf{A}(\mathbf{b})^{-1}/\partial b_k = -\mathbf{A}(\mathbf{b})^{-1}(\partial \mathbf{A}(\mathbf{b})/\partial b_k)\mathbf{A}(\mathbf{b})^{-1}$, where $\mathbf{A}(\mathbf{b}) = \sum b_k c_k \mathbf{x}_k \mathbf{x}_k^T$, we get

$$\begin{aligned} \partial(b_k g_k(\mathbf{b}))/\partial b_k &= g_k(\mathbf{b}) - \mathbf{x}_k^T \mathbf{A}(\mathbf{b})^{-1} b_k c_k \mathbf{x}_k \\ &\quad - (\mathbf{X} - \hat{\mathbf{X}}(\mathbf{b}))^T \mathbf{A}(\mathbf{b})^{-1} (c_k \mathbf{x}_k \mathbf{x}_k^T) \mathbf{A}(\mathbf{b})^{-1} (b_k c_k \mathbf{x}_k) \end{aligned} \quad (3.2)$$

and for $l \neq k$

$$\begin{aligned} \partial(b_l g_l(\mathbf{b}))/\partial b_k &= -\mathbf{x}_k^T \mathbf{A}(\mathbf{b})^{-1} (b_l c_l \mathbf{x}_l) \\ &\quad - (\mathbf{X} - \hat{\mathbf{X}}(\mathbf{b}))^T \mathbf{A}(\mathbf{b})^{-1} (c_k \mathbf{x}_k \mathbf{x}_k^T) \mathbf{A}(\mathbf{b})^{-1} (b_l c_l \mathbf{x}_l). \end{aligned} \quad (3.3)$$

It now follows from (3.2) and (3.3), that

$$\partial f(\mathbf{b})/\partial b_k = g_k(\mathbf{b}) e_k(\mathbf{b}), \quad (3.4)$$

where

$$e_k(\mathbf{b}) = y_k - \mathbf{x}_k^T \mathbf{B}(\mathbf{b}) \quad (3.5)$$

with $\mathbf{B}(\mathbf{b}) = \mathbf{A}^{-1}(\mathbf{b})(\sum b_k c_k \mathbf{x}_k \mathbf{x}_k^T)$. Therefore, $z_k = \partial f(\mathbf{b})/\partial b_k|_{\mathbf{b}=\mathbf{d}(s)}$ reduces to

$$z_k = g_k(\mathbf{d}(s)) e_k, \quad (3.6)$$

where $e_k = y_k - \mathbf{x}_k^T \hat{\mathbf{B}}$ with $\hat{\mathbf{B}} = \mathbf{B}(\mathbf{d}(s))$.

The variance estimator of \hat{Y}_w , resulting from (3.6), namely $v(z)$, takes account of the g -weights, $g_k(\mathbf{d}(s))$, unlike the standard linearization variance estimator (see e.g., Särndal *et al.* 1991, page 237). It agrees with the model-assisted variance estimator of Särndal *et al.* (1989). It also agrees with the jackknife linearization variance estimator when the latter is applicable (Yung and Rao 1996).

3.2 A General Class of Regression Calibration Weights

We now turn to a general class of regression calibration weights of the form $w_k(s) = d_k(s)h_k(\mathbf{d}(s))$ with

$$h_k(\mathbf{d}(s)) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{Q}}^{-1} (c_k \mathbf{x}_k + \sum_{l \neq k} d_l(s) c_l \mathbf{x}_l), \quad (3.7)$$

where the ab -th element of $\hat{\mathbf{Q}}$ is given by

$$\hat{q}_{ab} = \sum_{k=1}^N d_k(s) c_k \mathbf{x}_{ak} \mathbf{x}_{bk} + \sum_{k=1}^N \sum_{l \neq k}^N d_k(s) d_l(s) c_{kl} \mathbf{x}_{ak} \mathbf{x}_{bl}$$

for specified constants c_k and $c_{kl}(=c_{lk})$. The class (3.7) covers the GREG estimator as well as the “optimal” linear regression estimator with $d_k(s) = (1/\pi_k) a_k(s)$. In the former case $c_{kl} = 0$ while the optimal linear regression estimator uses $c_k = (1 - \pi_k)/\pi_k$ and $c_{kl} = (\pi_{kl} - \pi_k \pi_l)/\pi_{kl}$, $k \neq l$, where π_{kl} is the probability of including both elements k and l in the sample s (Montanari 1998).

The calibration weights $w_k(s)$ may be rewritten as

$$w_k(s) = d_k(s) + (X - \hat{X})^T \hat{Q}^{-1} \left(d_k(s) c_k \mathbf{x}_k + \sum_{l \neq k} d_{kl}(s) c_{kl}^* \mathbf{x}_l \right), \quad (3.8)$$

where $d_{kl}(s) = d_k(s) d_l(s) / E[d_k(s) d_l(s)]$, $c_{kl}^* = c_{kl} E[d_k(s) d_l(s)]$ and

$$\hat{Q}_{ab} = \sum_{k=1}^N d_k(s) c_k x_{ak} x_{bk} + \sum_{k=1}^N \sum_{l \neq k} d_{kl}(s) c_{kl}^* x_{ak} x_{bl}.$$

Note that $E d_k(s) = 1$ and $E d_{kl}(s) = 1$. If $d_k(s) = (1/\pi_k) a_k(s)$ then $d_k(s)$ reduces to $d_k(s) = a_k(s) a_l(s) / \pi_{kl}$ and $c_{kl} = (\pi_{kl} - \pi_k \pi_l) / (\pi_k \pi_l)$. We can regard the calibration estimator \hat{Y}_w resulting from (3.8) as a function of totals, by expressing a quadratic form as a total of synthetic variables (Sitter and Wu 2002). Therefore, we can use the method of section 2 and write $\hat{Y}_w = f(\mathbf{d}^{(1)}(s), \mathbf{d}^{(2)}(s), \mathbf{y}) = \sum d_k(s) h(\mathbf{d}^{(1)}(s), \mathbf{d}^{(2)}(s)) y_k$ where $\mathbf{d}^{(1)}(s) = \mathbf{d}(s)$ and $\mathbf{d}^{(2)}(s)$ is the vector of elements $d_{kl}(s)$, $k < l$, arranged in a sequence. Now, following the derivation of (2.3), we get

$$\hat{Y}_w - Y \approx \sum_k \tilde{z}_k (d_k(s) - 1) + 2 \sum_{k < l} \tilde{z}_{kl} (d_{kl}(s) - 1) \quad (3.9)$$

where

$$\tilde{z}_k = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_k |_{b^{(1)}=1, b^{(2)}=1},$$

$$\tilde{z}_{kl} = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_{kl} |_{b^{(1)}=1, b^{(2)}=1},$$

$\mathbf{b}^{(1)} = \mathbf{b} = (b_1, \dots, b_N)^T$ and $\mathbf{b}^{(2)}$ is the vector of arbitrary real numbers b_{kl} , $k < l$, arranged in the same order as the elements $d_{kl}(s)$ in $\mathbf{d}^{(2)}(s)$. Using (3.9), a variance estimator of \hat{Y}_w is approximately given by the variance estimator of $\sum_k \tilde{z}_k d_k(s) + 2 \sum_{k < l} \tilde{z}_{kl} d_{kl}(s)$, denoted by $v(\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)})$.

Since $v(\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)})$ involves the unknown values \tilde{z}_k and \tilde{z}_{kl} , we replace \tilde{z}_k by $\tilde{z}_k = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_k |_{b^{(1)}=\mathbf{d}^{(1)}(s), b^{(2)}=\mathbf{d}^{(2)}(s)}$ and \tilde{z}_{kl} by $\tilde{z}_{kl} = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_{kl} |_{b^{(1)}=\mathbf{d}^{(1)}(s), b^{(2)}=\mathbf{d}^{(2)}(s)}$ to get $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$. Unfortunately, the variance estimator $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ involves third order and fourth order moments $E[d_k(s) d_l(s) d_q(s)]$ and $E[d_k(s) d_l(s) d_q(s) d_r(s)]$ in addition to the second moments $E[d_k(s) d_l(s)]$, whereas the variance estimator for the generalized regression estimator requires only the second moments. In particular, if $d_k(s) = (1/\pi_k) a_k(s)$ we required third and fourth order inclusion probabilities $\pi_{k_l q}$ and $\pi_{k_l q r}$, as well as the second order inclusion probabilities π_{kl} .

The calculation of \tilde{z}_k and \tilde{z}_{kl} involves the derivatives $\partial [b_l h(\mathbf{b}^{(1)}, \mathbf{b}^{(2)})] / \partial b_k$ for $l = k$ and $l \neq k$ and the derivatives $\partial [b_l h(\mathbf{b}^{(1)}, \mathbf{b}^{(2)})] / \partial b_{kl}$ for $l = k$ and $l \neq k$. After simplification, we get

$$\tilde{z}_k = \left[1 + (X - \hat{X})^T \hat{Q}^{-1} c_k \mathbf{x}_k \right] e_k^*$$

and

$$\tilde{z}_{kl} = (X - \hat{X})^T \hat{Q}^{-1} c_{kl}^* \mathbf{x}_l e_k^*,$$

where

$$e_k^* = y_k - \mathbf{x}_k^T \hat{\mathbf{B}}^*$$

with $\hat{\mathbf{B}}^* = \hat{Q}^{-1} (\sum_k d_k(s) c_k \mathbf{x}_k y_k + \sum_{l \neq k} d_{kl}(s) d_l(s) c_{kl} \mathbf{x}_l y_k)$. Note that the customary Taylor linearization variance estimation uses $v(e^*)$, while $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ would involve the residuals e_k^* as well as the g -weights $1 + (X - \hat{X})^T \hat{Q}^{-1} c_k \mathbf{x}_k$ and $(X - \hat{X})^T \hat{Q}^{-1} c_{kl}^* \mathbf{x}_l$. If $c_{kl} = 0$ for all $k \neq l$, then $\tilde{z}_{kl} = 0$ and $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ reduces to $v(\mathbf{z})$ with \mathbf{z}_k given by (3.6). Thus the GREG result of subsection 3.1 is a special case.

3.3 Estimating Equations

We now turn to a vector parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ defined either explicitly or implicitly as the solution to “census” estimating equations $\mathbf{S}(\boldsymbol{\theta}) = \sum_{k=1}^N \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}$. A calibration estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^T$ with GREG calibration weights $w_k(s) = d_k(s) g_k(\mathbf{d}(s))$ is obtained as the solution to sample estimating equations:

$$\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}) = \sum w_k(s) \mathbf{u}_k(\hat{\boldsymbol{\theta}}) = \mathbf{0}, \quad (3.10)$$

where $\mathbf{u}_k(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}})$ are $(p \times 1)$ vectors (Binder 1983). For example for logistic regression with scalar θ , we have $u_k(\theta) = (y_k - p_k(\theta)) a_k$, where $p_k(\theta) = P(y_k = 1 | a_k) = \exp(\theta a_k) / (1 + \exp(\theta a_k))$ and a_k is the predictor variable. Note that $\hat{\theta}$, in this case, is the implicit solution to (3.10) and obtained iteratively using Newton-Raphson or Fisher scoring method.

The estimator of a ratio of totals Y and $A = \sum a_k$ is obtained as the explicit solution of (3.10) with $u_k(\theta) = y_k - \theta a_k$: $\hat{\theta} = \sum w_k(s) y_k / \sum w_k(s) a_k = \hat{Y} / \hat{A}$. In this case, $\hat{\theta}$ is a function of estimated totals and hence our method for functions of totals is applicable. It remains to evaluate $\partial f(\mathbf{b}) / \partial b_k$, where $f(\mathbf{b}) = \sum b_k g_k(\mathbf{b}) y_k / \sum b_k g_k(\mathbf{b}) a_k$. We have

$$\partial f(\mathbf{b}) / \partial b_k = \sum_{l=1}^N [\partial (b_l g_l(\mathbf{b})) / \partial b_k] \hat{A}(\mathbf{b})^{-1} (y_l - f(\mathbf{b}) a_l),$$

where $\hat{A}(\mathbf{b}) = \sum b_l g_l(\mathbf{b}) a_l$. Now using (3.4) and (3.5), it is easy to verify that \tilde{z}_k reduces to

$$\tilde{z}_k = g_k(\mathbf{d}(s)) \hat{A}^{-1} e_k^*$$

where

$$e_k^* = u_k(\hat{\theta}) - \mathbf{x}_k^T \hat{\mathbf{B}}_u$$

with $\hat{\mathbf{B}}_u$ obtained from $\hat{\mathbf{B}}$ by changing y_k to $u_k(\hat{\boldsymbol{\theta}})$. Note that the residuals e_k^* has the same form as the GREG residuals e_k with y_k changed with $u_k(\hat{\boldsymbol{\theta}})$.

In general, the solution $\hat{\boldsymbol{\theta}}$ to the estimating equations (3.10) may not be expressible as a function of estimated totals. We therefore follow Binder's (1983) approach and write the linearization estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ as

$$\mathbf{v}_L(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1} \hat{\boldsymbol{\Sigma}}_S(\hat{\boldsymbol{\theta}}) [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1}, \quad (3.11)$$

where $\hat{\mathbf{J}}(\boldsymbol{\theta}) = -\partial \hat{\mathbf{S}}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and $\hat{\boldsymbol{\Sigma}}_S(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix $\mathbf{v}_L(\hat{\mathbf{S}}(\boldsymbol{\theta})) = \hat{\boldsymbol{\Sigma}}_S(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Binder (1983) gave regularity conditions for the validity of (3.11). Noting that $\hat{\mathbf{S}}(\boldsymbol{\theta})$ is a vector of estimated totals with GREG weights $d_k(s) g_k(\mathbf{d}(s))$, it follows from (3.6) and (3.11) that

$$\mathbf{v}_L(\hat{\boldsymbol{\theta}}) = \mathbf{v}(\mathbf{z}) \quad (3.12)$$

where

$$\mathbf{z}_k = [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1} g_k(\mathbf{d}(s)) \mathbf{e}_k^* \quad (3.13)$$

with $\mathbf{e}_k^* = (e_{k1}^*, \dots, e_{kp}^*)^T$ and

$$e_{kj}^* = u_{jk}(\hat{\boldsymbol{\theta}}) - \mathbf{x}_k \hat{\mathbf{B}}_{ju}; \quad j = 1, \dots, p.$$

Further, $\hat{\mathbf{B}}_{ju}$ is obtained from $\hat{\mathbf{B}}_j$ by changing y_k to $u_{jk}(\hat{\boldsymbol{\theta}})$ and $\mathbf{v}(\mathbf{z})$ is the estimated covariance matrix of the vector of estimated totals $\hat{\mathbf{Z}} = \sum d_k(s) \mathbf{z}_k$, where $u_{jk}(\hat{\boldsymbol{\theta}})$ is the j^{th} element of $\mathbf{u}_k(\hat{\boldsymbol{\theta}})$. The result (3.12) agrees with the jackknife linearization variance estimator, \mathbf{v}_{JL} , for stratified multistage sampling obtained by Rao, Yung and Hidroglou (2002).

The result (3.12)–(3.13) may also be obtained directly by writing $\hat{\boldsymbol{\theta}}$ as $f(\mathbf{d}(s))$ and evaluating $\mathbf{z}_k = \partial f(\mathbf{b})/\partial b_k|_{\mathbf{b}=\mathbf{d}(s)}$. We denote $\hat{\boldsymbol{\theta}}(\mathbf{b}) = f(\mathbf{b})$ as the solution of $\sum (b_k g_k(\mathbf{b})) \mathbf{u}_k(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, i.e.,

$$\sum (b_k g_k(\mathbf{b})) \mathbf{u}_k(\hat{\boldsymbol{\theta}}(\mathbf{b})) = \mathbf{0}, \quad (3.14)$$

We now take the derivative of (3.14) with respect to b_k to get

$$\sum_{l=1}^N [\partial (b_l g_l(\mathbf{b})) / \partial b_k] \mathbf{u}_l(\hat{\boldsymbol{\theta}}(\mathbf{b})) + \sum_{l=1}^N (b_l g_l(\mathbf{b})) [\partial \mathbf{u}_l(\hat{\boldsymbol{\theta}}(\mathbf{b})) / \partial \hat{\boldsymbol{\theta}}(\mathbf{b}))] \partial \hat{\boldsymbol{\theta}}(\mathbf{b}) / \partial b_k. \quad (3.15)$$

Substituting (3.2) and (3.3) for $\partial (b_l g_l(\mathbf{b})) / \partial b_k$ in (3.15), we obtain (3.13) after simplification. This result shows that our method is also directly applicable to general estimators $\hat{\boldsymbol{\theta}}$ under Binder's (1983) regularity conditions.

3.4 A General Class of Calibration Estimators

The calibration weights, $w_k(s)$, associated with the GREG estimator \hat{Y}_w may not be always nonnegative. To get

around this difficulty, generalized raking ratio weights are often used. These weights are always nonnegative, but the method can lead to some extreme weights (Deville and Särndal 1992).

The generalized raking weights belong to the class

$$w_k(s) = d_k(s) F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \quad (3.16)$$

with $F(a) = e^a$, where the LaGrange multiplier $\hat{\boldsymbol{\lambda}}$ is determined by solving the calibration equations

$$\sum w_k(s) \mathbf{x}_k = \sum d_k(s) F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{x}_k = \mathbf{X}. \quad (3.17)$$

The GREG weights correspond to $F(a) = 1 + a$ in which case $\hat{\boldsymbol{\lambda}} = (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} (\mathbf{X} - \hat{\mathbf{X}})$.

In general, the calibration estimator $\hat{Y}_w = \sum w_k(s) y_k$ with weights $w_k(s)$ given by (3.16) may not be expressible as a function of estimated totals. We therefore follow Binder's (1983) approach and expand $F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ around $\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ denotes the probability limit of $\hat{\boldsymbol{\lambda}}$. We get

$$F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \approx F(\mathbf{x}_k^T \boldsymbol{\lambda}) + f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k^T (\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}), \quad (3.18)$$

where $f(a) = \partial F(a) / \partial a$. Further, by expanding the calibration equations (3.17) around $\boldsymbol{\lambda}$, we obtain after simplification,

$$\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \approx -\hat{\mathbf{Q}}_{\boldsymbol{\lambda}}^{-1} (\hat{\mathbf{S}}_{\boldsymbol{\lambda}} - \mathbf{X}) \quad (3.19)$$

where $\hat{\mathbf{Q}}_{\boldsymbol{\lambda}} = \sum d_k(s) f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^T$ and $\hat{\mathbf{S}}_{\boldsymbol{\lambda}} = \sum d_k(s) F(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k$. Note that both $\hat{\mathbf{Q}}_{\boldsymbol{\lambda}}$ and $\hat{\mathbf{S}}_{\boldsymbol{\lambda}}$ are of the form of estimated totals. Substituting (3.19) into (3.18) gives

$$F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \approx F(\mathbf{x}_k^T \boldsymbol{\lambda}) - f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k^T \hat{\mathbf{Q}}_{\boldsymbol{\lambda}}^{-1} (\hat{\mathbf{S}}_{\boldsymbol{\lambda}} - \mathbf{X}). \quad (3.20)$$

Using the approximation (3.20) in (3.16), it follows that \hat{Y}_w is approximated by a differentiable function of estimated totals. Hence, the general theory of section 2 is applicable and it remains to evaluate $\mathbf{z}_k = \partial h(\mathbf{b}) / \partial b_k|_{\mathbf{b}=\mathbf{d}(s)}$, where $h(\mathbf{b}) = \sum b_k g_k^*(\mathbf{b}) y_k$ with

$$g_k^*(\mathbf{b}) = F(\mathbf{x}_k^T \boldsymbol{\lambda}) - f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k^T \hat{\mathbf{Q}}_{\boldsymbol{\lambda}}^{-1} (\mathbf{S}_{\boldsymbol{\lambda}}(\mathbf{b}) - \mathbf{X})$$

where $\hat{\mathbf{Q}}_{\boldsymbol{\lambda}}(\mathbf{b}) = \sum b_k f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^T$ and $\mathbf{S}_{\boldsymbol{\lambda}}(\mathbf{b}) = \sum b_k F(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k$. After simplification, we get

$$\mathbf{z}_k = F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) (y_k - \mathbf{x}_k^T \hat{\mathbf{B}}_{\boldsymbol{\lambda}}) = F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) e_{k\lambda}, \quad (3.21)$$

where

$$\hat{\mathbf{B}}_{\boldsymbol{\lambda}} = \left(\sum d_k(s) f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum d_k(s) f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{x}_k y_k.$$

Singh and Folsom (2000) obtained a similar result, using a somewhat different approach.

The result (3.21) may also be obtained directly along the lines of (3.2) and (3.3) by writing \hat{Y}_w as $f(\mathbf{d}(s))$ and evaluating $\mathbf{z}_k = \partial f(\mathbf{b}) / \partial b_k|_{\mathbf{b}=\mathbf{d}(s)}$, where $f(\mathbf{b}) = \sum b_k g_k^*(\mathbf{b}) y_k$ with $g_k^*(\mathbf{b}) = F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b}))$. We have

$$\partial(b_k g_k(\mathbf{b}))/\partial b_k = g_k(\mathbf{b}) + b_k f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k^T (\partial \hat{\boldsymbol{\lambda}}(\mathbf{b})/\partial b_k), \quad (3.22)$$

and for $l \neq k$

$$\partial(b_l g_l(\mathbf{b}))/\partial b_k = b_l f(\mathbf{x}_l^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_l^T (\partial \hat{\boldsymbol{\lambda}}(\mathbf{b})/\partial b_k). \quad (3.23)$$

To evaluate $\partial \hat{\boldsymbol{\lambda}}(\mathbf{b})/\partial b_k$, we take the derivatives of the calibration equations (3.17) with $\mathbf{d}(s)$ replaced by \mathbf{b} : $\sum b_k F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k - \mathbf{X} = \mathbf{0}$. This gives

$$\mathbf{0} = F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k + \sum_l b_l f(\mathbf{x}_l^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_l \mathbf{x}_l^T (\partial \hat{\boldsymbol{\lambda}}(\mathbf{b})/\partial b_k)$$

or

$$\partial \hat{\boldsymbol{\lambda}}(\mathbf{b})/\partial b_k = -\left(\sum b_k f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k \mathbf{x}_k^T\right)^{-1} F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k. \quad (3.24)$$

Substituting (3.24) into (3.22) and (3.23), we get (3.21) after simplification.

Dewille and Särndal (1992) showed that the asymptotic variance of \hat{Y}_w for general $F(\cdot)$ is equivalent to the asymptotic variance of the GREG estimator which involves the “census” regression coefficient \mathbf{B} . Using this result they obtained a variance estimator of \hat{Y}_w for general $F(\cdot)$, by replacing \mathbf{B} by $\hat{\mathbf{B}} = (\sum w_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} \sum w_k(s) \mathbf{x}_k y_k$, where $w_k(s) = d_k(s) F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$. The resulting z_k agrees with our z_k given by (3.21) if $f(a) = F(a)$, i.e., in the case of generalized raking weights. In the case of GREG estimator, we have $F(x) = 1 + x$, $f(x) = 1$ and $\hat{\boldsymbol{\lambda}} = (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} (\mathbf{X} - \hat{\mathbf{X}})$. It readily follows that $F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ reduces to the customary $g - \text{weight}$ $g_k(\mathbf{d}(s)) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} \mathbf{x}_k$, and $e_{k\lambda} = y_k - \mathbf{x}_k^T \hat{\mathbf{B}}_{\lambda}$ reduces to $e_k = y_k - \mathbf{x}_k^T \hat{\mathbf{B}}$ with $\hat{\mathbf{B}} = (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} \sum d_k(s) \mathbf{x}_k y_k$. Note that our z_k in this case is different from the z_k of Dewille and Särndal (1992), but agrees with a commonly used z_k (Särndal, Swensson and Wretman 1989).

Our method, along the lines of section 3.3, can be extended to implicitly defined estimators, $\hat{\theta}_w$, obtained as solutions to estimating equations (3.10) based on the general calibration weights (3.16). Details are omitted for simplicity.

4. TWO-PHASE SAMPLING

We extend our method to two-phase sampling, assuming the estimator $\hat{\theta}$ of a parameter θ can be expressed as a differentiable function, $g(\hat{\mathbf{Y}}, \hat{\mathbf{X}}^{(1)})$, of estimated totals, $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_m)^T$, from the second-phase sample and estimated totals, $\hat{\mathbf{X}}^{(1)} = (\hat{X}_1^{(1)}, \dots, \hat{X}_p^{(1)})^T$, from the first-phase sample only. Here $\hat{Y}_i = \sum_{k=1}^N d_k(s) y_{ik}$, $i = 1, \dots, m$, $\hat{X}_j^{(1)} = \sum_{k=1}^N d_k^{(1)}(s_1) x_{jk}$, $j = 1, \dots, p$, $d_k^{(1)}(s_1)$ denotes the first-phase design weight attached to the k^{th} element with $d_k(s_1) = 0$ if k is not in the first-phase sample s_1 , and $d_k(s)$ is the final design weight attached to the k^{th} element with $d_k(s) = 0$ if k is not in the second-phase sample s . Further,

the parameter $\theta = g(\mathbf{Y}, \mathbf{X})$ with $\mathbf{Y} = (Y_1, \dots, Y_m)^T$ and $\mathbf{X} = (X_1, \dots, X_p)^T$ denoting the vectors of Y - and X - totals. For example, the two-phase ratio estimator, \hat{Y}_{R2} , is of the form $\hat{\theta} = g(\hat{\mathbf{Y}}, \hat{\mathbf{X}}, \hat{\mathbf{X}}^{(1)})$:

$$\begin{aligned} \hat{Y}_{R2} &= \frac{\hat{\mathbf{Y}}}{\hat{\mathbf{X}}} \hat{\mathbf{X}}^{(1)} = \hat{\mathbf{R}} \hat{\mathbf{X}}^{(1)} \\ &= \frac{\sum d_k(s) y_k}{\sum d_k(s) x_k} \left(\sum d_k^{(1)}(s_1) x_k \right). \end{aligned} \quad (4.1)$$

Note that $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2)^T$ with $\hat{Y}_1 = \hat{Y}$, $\hat{Y}_2 = \hat{\mathbf{X}}$, and $\hat{\mathbf{X}}^{(1)} = \hat{\mathbf{X}}^{(1)}$. Also, $\theta = g(\mathbf{Y}, \mathbf{X}, \mathbf{X}^{(1)}) = \mathbf{Y}$.

For simplicity, consider a $g(\cdot)$ such that $N^{-1}g(\cdot)$ tends to a limit. Taylor linearization of $\hat{\theta} = g(\hat{\mathbf{Y}}, \hat{\mathbf{X}}^{(1)})$ around (\mathbf{Y}, \mathbf{X}) gives

$$\begin{aligned} \hat{\theta} - \theta &= g(\hat{\mathbf{Y}}, \hat{\mathbf{X}}^{(1)}) - g(\mathbf{Y}, \mathbf{X}) \\ &\approx (\partial g(\mathbf{a}, \mathbf{a}^{(1)})/\partial \mathbf{a})^T|_{\mathbf{a}=\mathbf{Y}, \mathbf{a}^{(1)}=\mathbf{X}} (\hat{\mathbf{Y}} - \mathbf{Y}) \\ &\quad + (\partial g(\mathbf{a}, \mathbf{a}^{(1)})/\partial \mathbf{a}^{(1)})^T|_{\mathbf{a}=\mathbf{Y}, \mathbf{a}^{(1)}=\mathbf{X}} (\hat{\mathbf{X}}^{(1)} - \mathbf{X}). \end{aligned} \quad (4.2)$$

Let $\check{\mathbf{Y}} = \sum b_k \mathbf{y}_k$ and $\check{\mathbf{X}}^{(1)} = \sum b_k^{(1)} \mathbf{x}_k$ for arbitrary real numbers $\mathbf{b} = (b_1, \dots, b_N)^T$ and $\mathbf{b}^{(1)} = (b_1^{(1)}, \dots, b_N^{(1)})^T$. Also, let $g(\check{\mathbf{Y}}, \check{\mathbf{X}}^{(1)}) = f(\mathbf{b}, \mathbf{b}^{(1)}, \mathbf{A}_y, \mathbf{A}_x) = f(\mathbf{b}, \mathbf{b}^{(1)})$, where \mathbf{A}_y is an $m \times N$ matrix with k^{th} column $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^T$, $k = 1, \dots, N$, and \mathbf{A}_x is a $p \times N$ matrix with k^{th} column $\mathbf{y}_k = (y_{k1}, \dots, y_{km})^T$, $k = 1, \dots, N$. Now following the derivation of (2.3) and noting that $\hat{\mathbf{Y}} = \mathbf{A}_y \mathbf{d}(s)$, $\mathbf{Y} = \mathbf{A}_y \mathbf{1}$, $\hat{\mathbf{X}}^{(1)} = \mathbf{A}_x \mathbf{d}^{(1)}(s_1)$, $\mathbf{X} = \mathbf{A}_x \mathbf{1}$, it can be shown that (4.2) reduces to

$$\hat{\theta} - \theta \approx \tilde{\mathbf{z}}^T (\mathbf{d}(s) - \mathbf{1}) + \tilde{\mathbf{z}}^{(1)T} (\mathbf{d}^{(1)}(s_1) - \mathbf{1}), \quad (4.3)$$

where $\mathbf{d}(s) = (d_1(s), \dots, d_N(s))^T$ and $\mathbf{d}^{(1)}(s_1) = (d_1^{(1)}(s_1), \dots, d_N^{(1)}(s_1))^T$. Further, $\tilde{\mathbf{z}} = (\tilde{z}_1, \dots, \tilde{z}_m)^T$ with $\tilde{z}_k = \partial f(\mathbf{b}, \mathbf{b}^{(1)})/\partial b_k|_{\mathbf{b}=\mathbf{1}, \mathbf{b}^{(1)}=\mathbf{1}}$, and $\tilde{\mathbf{z}}^{(1)} = (\tilde{z}_1^{(1)}, \dots, \tilde{z}_N^{(1)})^T$ with $\tilde{z}_k^{(1)} = \partial f(\mathbf{b}, \mathbf{b}^{(1)})/\partial b_k^{(1)}|_{\mathbf{b}=\mathbf{1}, \mathbf{b}^{(1)}=\mathbf{1}}$. It follows from (4.3) that a variance estimator of $\hat{\theta}$ is approximately given by the variance estimator of the estimated total $\sum d_k(s) \tilde{z}_k + \sum d_k^{(1)}(s_1) \tilde{z}_k^{(1)} = \hat{\mathbf{Y}}(\tilde{\mathbf{z}}) + \hat{\mathbf{X}}^{(1)}(\tilde{\mathbf{z}}^{(1)})$. We denote the latter variance estimator as $v(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}^{(1)})$. Now we replace \tilde{z}_k and $\tilde{z}_k^{(1)}$ by $z_k = \partial f(\mathbf{b}, \mathbf{b}^{(1)})/\partial b_k|_{\mathbf{b}=\mathbf{d}(s), \mathbf{b}^{(1)}=\mathbf{d}^{(1)}(s_1)}$ and $z_k^{(1)} = \partial f(\mathbf{b}, \mathbf{b}^{(1)})/\partial b_k^{(1)}|_{\mathbf{b}=\mathbf{d}(s), \mathbf{b}^{(1)}=\mathbf{d}^{(1)}(s_1)}$ respectively, since \tilde{z}_k and $\tilde{z}_k^{(1)}$ are unknown. This leads to a linearization variance estimator

$$v_L(\hat{\theta}) = v(\mathbf{z}, \mathbf{z}^{(1)}). \quad (4.4)$$

We now consider the special case of a “double expansion” estimator $\hat{Y}(y) = \sum d_k(s) y_k$ with $d_k(s) = \pi_{1k}^{-1} \pi_{2k/1}^{-1}$ for $k \in s$ and the Horvitz-Thompson (H-T) estimator $\hat{X}^{(1)}(x) = \sum d_k^{(1)}(s_1) x_k$ with $d_k^{(1)}(s_1) = \pi_{1k}^{-1}$ for $k \in s_1$, where π_{1k} is the probability of including element k in s_1 , and $\pi_{2k/1}$ is the conditional probability of including element k in s

given s_1 . In this case, an unbiased H-T type estimator of $\hat{Y}(y) + \hat{X}^{(1)}(x)$ is given by

$$\begin{aligned} v(y, x) = & \sum_{k,l \in s_1} \sum \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{1kl}} \frac{x_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}} \\ & + \sum_{k,l \in s} \sum \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{kl}^*} \left(\frac{y_k}{\pi_{1k}} \frac{y_l}{\pi_{1l}} + 2 \frac{y_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}} \right) \\ & + \sum_{k,l \in s} \sum \frac{\pi_{2kl/l} - \pi_{2k/l} \pi_{2l/l}}{\pi_{2kl/l}} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*} \end{aligned} \quad (4.5)$$

where $\pi_k^* = \pi_{1k} \pi_{2k/l}$, $\pi_{kl}^* = \pi_{1kl} \pi_{2kl/l}$, π_{1kl} is the probability of including both elements k and l in s_1 and $\pi_{2kl/l}$ is the conditional probability of including both elements k and l in s given s_1 . A proof of (4.5) is given in the Appendix. The variance estimator (4.4) is obtained from (4.5) by changing y_k and x_k to z_k and $z_k^{(1)}$ respectively.

Example 4.1 We illustrate the calculation of $v(z, z^{(1)})$ for the two-phase ratio estimator \hat{Y}_{R2} , given by (4.1), for the special case of simple random sampling at both phases: s_1 is a simple random sample of size n and s is a simple random subsample of size m from s_1 . In this case, $\pi_{1k} = n/N$ and $\pi_{2k/l} = m/n$. Further, it follows from (4.1) that for general two-phase design,

$$z_k = \frac{\hat{X}^{(1)}}{\hat{X}} (y_k - \hat{R}x_k) = \frac{\hat{X}^{(1)}}{\hat{X}} e_k \quad (4.6)$$

and

$$z_k^{(1)} = \hat{R} x_k. \quad (4.7)$$

Under simple random sampling at both stages, (4.6) and (4.7) reduce to $z_k = (\bar{x}^{(1)}/\bar{x}) e_k$ and $z_k^{(1)} = (\bar{y}/\bar{x}) x_k$, where $e_k = y_k - (\bar{y}/\bar{x}) x_k$, \bar{y} and \bar{x} are the second-phase sample means of y and x respectively, and $\bar{x}^{(1)}$ is the first-phase sample mean of x . Now substituting z_k and $z_k^{(1)}$ for y and x in (4.5) and noting that $\pi_{1kl} = n(n-1)/[N(N-1)]$, $\pi_{2kl/l} = m(m-1)/[n(n-1)]$, $\pi_{1kk} = \pi_{1k}$ and $\pi_{2kk/l} = \pi_{2k/l}$, we get

$$\begin{aligned} v_L(\hat{Y}_{R2}) = & N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \hat{R}^2 s_{1x}^2 + N^2 \left(\frac{1}{m} - \frac{1}{n} \right) \left(\frac{\bar{x}^{(1)}}{\bar{x}} \right)^2 s_{2e}^2 \\ & + 2N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \hat{R} \frac{\bar{x}^{(1)}}{\bar{x}} s_{ex}, \end{aligned} \quad (4.8)$$

where

$$\hat{R} = \bar{y}/\bar{x}, \quad s_{1x}^2 = (n-1)^{-1} \sum_{k \in s_1} (x_k - \bar{x}^{(1)})^2,$$

$$s_{2e}^2 = (m-1)^{-1} \sum_{k \in s} (e_k - \bar{e})^2,$$

$$s_{2ex} = (m-1)^{-1} \sum_{k \in s} (e_k - \bar{e})(x_k - \bar{x})$$

and \bar{e} is the second-phase sample mean of e . The formula (4.8) agrees with the formula derived by Rao and Sitter (1995). It is different from the customary formula (Sukhatme and Sukhatme 1970, page 176) which fails to make use of the full x -data $\{x_k, k \in s_1\}$. Rao and Sitter (1995) demonstrated through simulation that $v_L(\hat{Y}_{R2})$ is more efficient than the customary variance estimator. Also, $v_L(\hat{Y}_{R2})$ performed better in tracking the conditional mean squared error of \hat{Y}_{R2} ; see Rao and Sitter (1995, section 3) for details of the simulation study.

CONCLUDING REMARKS

We have presented a unified approach to deriving Taylor linearization variance estimators and applied it to a variety of problems. It leads directly to a variance estimator that has some desirable properties at least in a number of important special cases; in particular, approximate unbiasedness for the model variance of the estimator under an assumed model and validity under a conditional repeated sampling framework. It would be useful to investigate whether such desirable properties also hold for more complex cases such as the general class of calibration estimators (section 3.2), the estimators based on estimating equations (section 3.3) and two-phase sampling (section 4). We are currently investigating various extensions of our method, including variance estimation under imputation for item nonresponse and variance estimation from longitudinal survey data.

ACKNOWLEDGMENTS

We thank the Associate Editor and a referee for constructive comments and suggestions. We also thank several colleagues in Statistics Canada for useful suggestions and encouragement, especially Linda Standish, David Binder, Geoff Hole, Richard Burgess and Larry Swain. Demnati's work was made possible by the Small Area and Administrative Data Division of Statistics Canada. J.N.K. Rao's work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

APPENDIX

Unbiased Variance Estimator of $\hat{Y}(y) + \hat{X}^{(1)}(x)$

The variance of $\hat{Y}(y) + \hat{X}^{(1)}(x)$ is the sum of the variance of $\hat{Y}(y)$, the variance of $\hat{X}^{(1)}(x)$ and twice the covariance of $\hat{Y}(y)$ and $\hat{X}^{(1)}(x)$. An unbiased H-T type estimator of

$V[\hat{Y}(y)]$ is given by Särndal, Swensson and Wretman (1991, chapter 9, page 348):

$$v[\hat{Y}(y)] = \sum_{k,l \in s} \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{kl}^*} \frac{y_k}{\pi_{1k}} \frac{y_l}{\pi_{1l}} + \sum_{k,l \in s} \frac{\pi_{2kl/1} - \pi_{2k/1} \pi_{2l/1}}{\pi_{2kl/1}} \frac{y_k}{\pi_{2k}^*} \frac{y_l}{\pi_{2l}^*}. \quad (A.1)$$

An unbiased H-T type estimator of $V[\hat{X}^{(1)}(x)]$ is given by

$$v[\hat{X}^{(1)}(x)] = \sum_{k,l \in s_1} \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{1kl}} \frac{x_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}}. \quad (A.2)$$

Further,

$$\text{Cov}[\hat{Y}(y), \hat{X}^{(1)}(x)] = E\text{Cov}_2[\hat{Y}(y), \hat{X}^{(1)}(x)] + \text{Cov}[E_2(\hat{Y}(y)), E_2(\hat{X}^{(1)}(x))],$$

where E_2 and Cov_2 denote conditional expectation and conditional covariance given s_1 . Noting that

$$E_2 \hat{Y}(y) = \hat{X}^{(1)}(y), E_2 \hat{X}^{(1)}(x) = \hat{X}^{(1)}(x)$$

and $\text{Cov}_2[\hat{Y}(y), \hat{X}^{(1)}(x)] = 0$, we get

$$\text{Cov}[\hat{Y}(y), \hat{X}^{(1)}(x)] = \text{Cov}[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)].$$

An unbiased H-T type estimator of $2\text{Cov}[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)]$ is given by

$$2\text{cov}[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)] = 2 \sum_{k,l \in s} \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{kl}^*} \frac{y_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}}. \quad (A.3)$$

The sum of (A.1), (A.2) and (A.3) equals (4.5).

REFERENCES

- ANDERSON, C., and NORDBERG, L. (1994). A method for variance estimation of non-linear functions of totals in surveys - theory and software implementation. *Journal of Official Statistics*. 10, 395-405
- BERGER, Y.G. (2002). A generalized jackknife variance estimator for nonlinear statistics in probability sampling. Technical Report, Department of Social Statistics, University of Southampton.
- BINDER, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. 51, 279-292.
- BINDER, D. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*. 22, 17-22.
- CAMPBELL, C. (1980). A different view of finite population estimation. *Proceeding of the Section on Survey Research Methods*, American Statistical Association. 319-324.
- DEVILLE, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*. 25, 193-203.
- DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87, 376-382.
- HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons, Inc.
- HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons, Inc.
- KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*. 9, 1010-1019.
- MONTANARI, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*. 24, 69-77.
- RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*. 82, 453-460.
- RAO, J.N.K., YUNG, W. and HIDIROGLOU, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*. 76, 66-77.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*. 76, 527-537.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J.H. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SERFLING, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons, Inc.
- SINGH, A.C., and FOLSOM, R.E. (2000). Bias correcting estimating function approach for variance estimation adjusted for poststratification. *Proceeding of the Section on Survey Research Methods*, American Statistical Association. 610-615.
- SITTER, R.R., and WU, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association*. 97, 535-544.
- SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. 2nd ed. London: Asia Publishing House.
- VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*. 88, 89-96.
- YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*. 22, 23-31.

Comment

PHILLIP S. KOTT¹

The article addresses an impressive number of contexts, many of which have only recently been investigated in the literature, often by Professor Rao himself. I will have little to say here about estimating functions with calibration weights or two-phase sampling, except (mostly) to agree with the solutions advocated in the text. Instead, I will focus on three applications: the ratio estimator under simple random sampling discussed in the Introduction, the general class of regression calibration weights from section 3.2, and the general class of calibration estimators from section 3.4. I will end with a question about the linearization variance estimator in full Horvitz-Thompson form, which has bothered me for some time.

The Ratio Under Simple Random Sampling

Before beginning, let me confess to a certain skepticism about the general method proposed in section 2. I find that techniques of this sort work best when you already know what the answer is. Godambe and Thompson (1986) tried to use estimating functions to settle a controversy then surrounding the best variance estimator for the ratio under simple random sampling. Using the notation in the text, they demonstrated that $(\bar{X}/\bar{x})v_L$ was the proper way to estimate the variance of a ratio estimator, $\hat{Y}_R = (\bar{X}/\bar{x})\bar{y}$. Later, Binder (1996) corrected them. He showed that when done properly, $v_{JL} = (\bar{X}/\bar{x})^2 v_L$ is produced from estimating-function technology. It helped that he already knew that was the better answer.

As Demanti and Rao state, v_{JL} has both good randomization (design) and model-based properties (here and hereafter I omit the qualifier, "under mild conditions which I assume to hold"). In fact, when n/N is ignorably small, v_{JL} has a relative bias of $O(1/n)$ as an estimator for the model variance of \hat{Y}_R . If the y_k are uncorrelated, then this is not only true when $V_m(y_k) = \sigma^2 x_k$ as stated in the text, but, more generally, when $V_m(y_k) = \sigma_k^2$. Unfortunately, the result is less general when n/N is not ignorably small. In that context, when the y_k are uncorrelated and $V_m(y_k) = \sigma^2 x_k$, a more appropriate estimator for the model variance of \hat{Y}_R is $v_m = [(\bar{X}/\bar{x})^2 - (n/N)(\bar{X}/\bar{x})][1 - (n/N)]^{-1} v_L$ (Kott and Brewer 2001). As an estimator for the randomization mean squared error of \hat{Y}_R , v_m has a relative bias of $O(1/\sqrt{n})$, just like v_{JL} and v_L .

When simple random sampling is used in practice the sampling fraction is almost always small. Thus, v_{JL} is an

attractive variance/mean-squared-error estimator, and my criticism of Demnati and Rao for advocating it is mild.

A General Class of Regression Calibration Weights

I would generalize the results of section 3.1 in a different manner than the authors do in section 3.2. Following Estavao and Särndal (2002), replace $c_k x_k$ in equation (3.1) with a vector q_k having the same dimension as x_k . The rest of that section follows easily.

One choice for q_k is

$$q_{(1)k} = \sum_{j \in U} (\pi_{kj} - \pi_k \pi_j) x_j / (\pi_k \pi_j),$$

the use of which results in a variant of the randomization-optimal regression estimator proposed by Tillé (1999). Observe that $(\sum_U q_{(1)k} x_k^T)^{-1} (\sum_U q_{(1)k} y_k^T) = [\text{Var}(\hat{X})]^{-1} \text{Cov}(\hat{X}, \hat{Y})$, where Var and Cov denote randomization-based properties.

Another choice, investigated indirectly by Demnati and Rao and likewise resulting into a variant of the randomization-optimal estimator, is

$$q_{(2)k} = \sum_{j \in s} (\pi_{kj} - \pi_k \pi_j) x_j / (\pi_k \pi_j).$$

Since $q_{(2)k}$ is a function of the sample, the authors take us through the complications of section 3.2. This was only necessary for randomization-based inference. I would have gone a different way. Observe that $d_k(s)q_{(2)k} - d_k(s)q_{(1)k} = O_p(1/\sqrt{n})$. Replacing one for the other has an asymptotically ignorable effect on $w_k(s)$ (i.e., the relative difference is $O_p(1/n)$).

A General Class of Calibration Estimators

A mild generalization of equation (3.16) allows calibration weights of the form,

$$w_k(s) = d_k(s)F(q_k^T \hat{\lambda}),$$

where q_k again has the same dimension as x_k . For convenience F is assumed positive and twice differentiable around $q_k^T \hat{\lambda}$. Without loss of generality, one can assume $\hat{\lambda}$ (the limit of $\hat{\lambda}$) is $\mathbf{0}$, and $f(0) > 1$. When $\hat{Y}_{GC} = \sum_U w_k(s)y_k$ is a randomization consistent estimator, as I assume it is, $F(0)$ is equal to 1.

Paralleling the development in the text leads ultimately to

$$z_k = F(q_k^T \hat{\lambda})(y_k - x_k^T \hat{B}_\lambda) = F(q_k^T \hat{\lambda})e_{k\lambda},$$

¹ Phillip S.Kott, USDA / NASS, 3251 Old Lee Hwy, Fairfax, VA 22030, U.S.A.

where $\hat{\mathbf{B}}_\lambda = [\sum d_k(s) f(\mathbf{q}_k^T \hat{\lambda}) \mathbf{q}_k \mathbf{q}_k^T]^{-1} \sum d_k(s) f(\mathbf{q}_k^T \hat{\lambda}) \mathbf{q}_k y_k$. The presence of the $f(\cdot)$ in the expression of $\hat{\mathbf{B}}_\lambda$ may be a bit of a surprise, but, it turns out, not a meaningful one in this context. For inference under the prediction model, $E_m(y_k | \mathbf{x}_k) = \mathbf{x}_k^T \boldsymbol{\beta}$, the derivative can be replaced by any constant without asymptotic consequence; $\hat{\mathbf{B}}_\lambda$ remains a model unbiased estimator for $\boldsymbol{\beta}$. For randomization-based inference, since $\mathbf{q}_k^T \hat{\lambda} = O_p(1/\sqrt{n})$ and $F(0), f(0) > 0$, z_k would be unaffected asymptotically if $f(\mathbf{q}_k^T \hat{\lambda})$ were replaced by 1 or by $F(\mathbf{q}_k^T \hat{\lambda})$.

Things change, however, if we push the envelop a bit. Fuller, Loughin and Baker (1994) use calibration to adjust for unit nonresponse by treating sample response as a second phase of sampling. They assume that every element k in the population has a Poisson probability of sample response, π_{2k} , which is independent of whether it is actually chosen for the sample. They further assume $\pi_{2k} = 1/(1 + \mathbf{x}_k^T \hat{\lambda})$, where $\hat{\lambda}$ is unknown and implicitly estimated by calibration. Here we generalize that and assume $\pi_{2k} = 1/F(\mathbf{q}_k^T \hat{\lambda})$, where F is known, positive, and twice differentiable. In practice, \mathbf{q}_k will likely be identical to \mathbf{x}_k , but it may be reasonable to replace one of more components of \mathbf{x}_k with variables conjectured to be more strongly correlated with response/nonresponse.

Redefining s as the respondent sample and $d_k(s)$ as $(1/\pi_{1k})$ when $k \in s, 0$ otherwise, everything proceeds as before. The difference is that $f(\mathbf{q}_k^T \hat{\lambda})$ in $\hat{\mathbf{B}}_\lambda$ need no longer need be asymptotically identical across the k . Thus, the term can matter even with a large sample.

Now $V(\hat{Y}_{GC}) \approx V(\sum_U d_k(s) z_k)$, where $\sum_U d_k(s) z_k = \sum_U d_k(s) F(\mathbf{q}_k^T \hat{\lambda}) e_{k\lambda}$ is the double expansion estimation. Substituting $1/F(\mathbf{q}_k^T \hat{\lambda})$ for π_{2k} , the variance estimator for \hat{Y}_{GC} becomes (from equation (A.1) with $\pi_{2kj/1} = \pi_{2kj} \pi_{2j}$)

$$\begin{aligned} v(\hat{Y}_{GC}) &= \sum_{k,j \in s} [(\pi_{1kj} - \pi_{1k} \pi_{1j}) / \pi_{1kj}] \\ &\quad d_k(s) F(\mathbf{q}_k^T \hat{\lambda}) e_{k\lambda} d_j(s) F(\mathbf{q}_j^T \hat{\lambda}) e_{j\lambda} \\ &\quad + \sum_{k \in s} \pi_{1k} \{ [F(\mathbf{q}_k^T \hat{\lambda})]^2 - [F(\mathbf{q}_k^T \hat{\lambda})] \} [d_k(s) e_{k\lambda}]^2. \end{aligned}$$

This differs from the variance estimator in Folsom and Singh (2000) mainly because those authors assume the original sample is chosen using a stratified multistage design employing with-replacement sampling in the first. That, among other things, annihilates the second summation on the right hand side.

Not only does $v(\hat{Y}_{GC})$ estimate the quasi-randomization mean squared error of \hat{Y}_{GC} —“quasi” because a response model is assumed, it also estimates the model variance of \hat{Y}_{GC} . In fact, the relative bias of $v(\hat{Y}_{GC})$ under the prediction model, $E_m(y_k | \mathbf{x}_k, \mathbf{q}_k) = \mathbf{x}_k^T \boldsymbol{\beta}$, is $O(1/n)$ when the y_k

are uncorrelated and $V_m(y_k | \mathbf{x}_k, \mathbf{q}_k) = \mathbf{x}_k^T \boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ (like $\boldsymbol{\beta}$) need not be specified. Surprisingly, the second term in $v(\hat{Y}_{GC})$ provides the model-based correction I recommended for the ratio estimator under simple random sampling in the absence of nonresponse.

Does the “Plug-in” Variance Estimator Really Work for the Full Horvitz-Thompson Form?

As I warned parenthetically early on, I have omitted the key phrase, “under mild conditions which I assume to hold,” repeatedly in these comments. Now, I want to turn my attention to what may be one of those conditions. It is standard in variance estimation to replace population (or model) values with sample analogues since their difference is asymptotically ignorable. That is done, for example, by Demnati and Rao in equation (2.4) when they plug in z_k for \tilde{z}_k . The question I want to raise, and for which I do not know the answer, is this. Suppose one is estimating a total with a calibration estimator. The total is $O(N)$, and $O(n) = O(N)$. The estimator’s model variance and randomization mean squared error are also $O(n)$. Is it legitimate to plug in z_k for \tilde{z}_k , where $z_k - \tilde{z}_k = O_p(1/\sqrt{n})$, when there are $n(n-1)/2$ terms in the Horvitz-Thompson – or Yates-Grundy – variance/mean-squared-error estimator? In most practical applications, this is a non-issue, because the variance estimator can be re-expressed with $O(n)$ terms. What if that is not the case?

Let me conclude these remarks by thanking Drs. Demnati and Rao for their stimulating article and *Survey Methodology* for both publishing it and allowing me to provide some comments.

ADDITIONAL REFERENCES

- ESTEVAO, V.M., and SÄRNDAL, C-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*. 18, 233-255.
- FULLER, W.A., LOUGHIN, M.M. and BAKER, H.D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology*. 20, 75-85.
- GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *International Statistical Review*. 54, 2, 127-138.
- KOTT, P.S., and BREWER K.R.W. (2001). Estimating the model variance of a randomization-consistent regression estimator. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 811-822.
- TILLÉ, Y. (1999). Estimation in surveys using conditional inclusion probabilities: complex designs. *Survey Methodology*. 25, 57-66.

Comment

BABUBHAI V. SHAH¹

This is an excellent paper that removes the mystery underlying Taylor linearization. Most data analysis applications use Horvitz-Thompson weights that are reciprocals of the probabilities of selection. The simplest prescription for deriving the linearization for an estimator $\hat{\theta}$ is as follows:

1. For each observation, create a new variable $z_i = \partial \hat{\theta} / \partial w_i$, where w_i is the reciprocal of the selection probability for the i -th observation selected in the sample. In cases where the estimator $\hat{\theta}$ is defined implicitly through estimating equations, the derivative can be computed by differentiating the implicit equations.
2. Define weighted $\hat{T} = \sum w_i z_i$ total.
3. Compute the variance \hat{V} of the total \hat{T} based on the sample design.
4. The variance \hat{V} is the approximate variance of the estimator $\hat{\theta}$.

If the parameter θ is a vector then the variable z_i and the total T are also vectors and \hat{V} is an approximate estimate of the variance covariance matrix of the estimator $\hat{\theta}$.

The steps (1) and (2) specified above produce the correct linearization in the following cases:

- a. Means, proportions, and ratio estimates.
- b. Generalized linear regression models.
- c. Predicted marginal for generalized linear model.
- d. Estimate of the mean from regression imputed data.
- e. Generalized linear regression models with calibrated weights.
- f. Wilcoxon two sample rank sum test.
- g. Estimates of coefficients and the hazard rate in Cox's proportional hazard model.
- h. Estimates of predicted marginal survival in Cox's proportional hazard model.
- i. Two-phase sample survey.

The derivation in the step (1) is uniquely defined and does not contain the true value of the parameter θ , and does not require substitution by the estimator $\hat{\theta}$.

The independence of step (3) for variance computation from the linearization in steps (1) and (2) is aptly demonstrated by the discussion on two-phase sampling in section 4. In most cases, one assumes with replacement sample design to estimate the variance of the total in the step (3). Of course, a better estimate of the variance of the total may be obtained by using all the available information about the sample design. For the case of a two-phase design, step (1) can be performed by using Horvitz Thompson weights for the phase one sampling, and treating the multipliers m_i as data. The multiplier m_i is equal to zero if the observation i is not selected in phase two and is equal to the inverse of the conditional probability $\pi_{2k/1}^{-1}$. The resulting step (2) produces the same total as presented in the paragraph between equations (4.3) and (4.4). The subsequent discussion in section 4, describes the appropriate way to estimate the variance of this total for a two-stage sample design without replacement at each stage, and that calculation is independent of the linearization.

The steps (1) and (2) generate appropriate linearization in all known cases except where the estimator is not a continuous function of the weights w_i , e.g., quantile.

¹ Babubhai V. Shah, SAFAL Institute, Inc. E-mail: babushah@earthlink.net

Comment

CHRIS SKINNER¹

Linearization and replication approaches provide two broad classes of methods for variance estimation in surveys. Both have their relative advantages and it seems important to keep a place for both in the survey statistician's 'toolkit'. This paper deepens our understanding of linearization methods, proposes a general procedure to generate such variance estimators uniquely and provides valuable illustrations of this procedure in some important areas of application.

A linearization method approximates the variance of a statistic of interest by the variance of a linear statistic, for which it is assumed a suitable variance estimator is available. The main issue here is the method used to determine the linear statistic. The standard approach assumes the statistic of interest may be expressed as a differentiable function of a vector of linear statistics (of fixed dimension) and uses Taylor series expansion to determine the approximation. The approach proposed in this paper applies to a more general class of sample-weighted statistics, illustrated by the complex examples in sections 3.2. and 4. The variance estimator is constructed by differentiating the statistic with respect to the sample weights. The approach to linear approximation is closely related to methods based upon the influence function (e.g., equations 1.6 and 1.13) and the paper provides a helpful review of such methods in section 1. The authors note that it is not easy to verify the validity of such methods for statistics which are not smooth functions of (or a fixed number of) linear statistics and it would be interesting to know how far the proposed approach does indeed provide valid variance estimators for statistics, such as quantiles, which are not of this form.

A key feature of the proposed approach, which ensures the unique construction of the variance estimator, is that derivatives are evaluated at values based on the achieved sample, without any initial evaluation of the approximating linear statistic at theoretical population values. Such initial evaluation may lead to non-uniqueness when auxiliary information is available, for example on a population mean, \bar{X} , and it is assumed that this value is equal to the limiting value of a corresponding sample statistic, \bar{x} . For statistics which are smooth functions of linear statistics, it appears that the variance estimator generated by the proposed method may also be constructed by conventional Taylor series methods, provided no initial simplification of the

variance estimator takes place based on such assumptions about auxiliary information. Such construction may, however, be less clear-cut than for the proposed approach.

Assumptions employed by linearization methods differing from the proposed approach, such as that an auxiliary value \bar{X} is the theoretical limiting value of a sample value \bar{x} , are based upon unconditional distributions and so it might be anticipated that the incorporation of such assumptions into a variance estimator might damage the method's conditional properties, especially with respect to statistics such as \bar{x} . The proposed procedure avoids dependence upon such assumptions and, by evaluating derivatives at achieved sample values, may be expected to track conditional properties more closely. (There appear to be parallels with Efron and Hinkley's (1978) arguments in favour of the observed versus the expected information, although the context is rather different.)

The avoidance of dependence upon such assumptions may not only benefit the conditional properties of the proposed approach, but also protect the variance estimator against possible biasing effects of non-sampling errors. The auxiliary population information may differ from the limiting values of the corresponding sample statistics either because of non-response or non-coverage or because of discrepancies in the way the auxiliary variables are measured. In such circumstances, linearization methods differing from the proposed approach might lead to inconsistent variance estimation. For this reason, Fuller (2002, page 10) recommends the use of the g -weights in (3.6), as proposed, especially in the presence of nonresponse (page 15). With regards to the latter case, it seems worth noting that the validity of the proposed procedure does not appear to depend on the requirement that $E(\mathbf{d}(s)) = \mathbf{1}$, provided $\mathbf{1}$ is replaced by $E(\mathbf{d}(s))$ in the development in section 2. In particular, if s denotes unit respondents and non-response may be represented by Poisson sampling with unknown response probabilities then the proposed approach to variance estimation may still be consistent (when based on many standard variance estimators for linear statistics), even if $d(s)$ is based only on sampling inclusion probabilities.

Julia d'Arrigo and I have recently studied the properties of linearization variance estimators under nonresponse in simulation studies as part of the DACSEIS research project (www.dacseis.de) using data from the UK Labour Force

¹ Chris Skinner, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, United Kingdom.
E-mail: cjs@socsci.soton.ac.uk.

Survey and the German Income and Expenditure Survey. We considered various calibration estimators under Poisson models for unit non-response which were ignorable given the calibrating variables, using standard variance estimators for linear statistics under stratified multi-stage sampling. We indeed found that nonresponse could lead to serious biases in the linearization variance estimators if they failed to take account of the g -weights for GREG estimation (section 3.1.) or ignored the $F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ term in (3.21). Such biases were absent in the proposed approach.

We also investigated the alternative calibration estimators discussed in section 3.4. Deville and Särndal's (1992) theoretical finding that the asymptotic variance of \hat{Y}_w does not depend on the form of the function $F(\cdot)$ is based on the assumption that $\sum d_k(s) \mathbf{x}_k$ is consistent for \mathbf{X} . This assumption may not hold under various sources of non-sampling error, and is not required for the proposed approach. Hence, the appropriate approximate linear statistic (under departures from this assumption) is defined by (3.21) and the resulting variance estimator may depend on the form of $F(\cdot)$, even asymptotically. The standard linearization variance estimators in which $d_k(s)f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ in $\hat{\mathbf{B}}_{\lambda}$ is replaced by $d_k(s)$ or $w_k(s)$ may be inconsistent if these weights differ from $d_k(s)f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$. Despite this theoretical fact, we observed little difference in our simulation study (for each of the functions, $1+u$, $\exp(u)$, and $(1-u)^{-1}$, used for $F(u)$) between the statistical properties of variance estimators based upon these three different choices of weight, $d_k(s)f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$, $d_k(s)$ or $w_k(s)$, in the $\hat{\mathbf{B}}_{\lambda}$ vector in (3.21). Others studies might produce different findings.

A disadvantage of the linearization methods considered here compared to replication methods is the need for analytic differentiation. It would appear from the examples presented in this paper that the analytic differentiation involved in the proposed method is at least as straightforward as that in standard methods of Taylor series expansion of smooth functions of linear statistics. Nevertheless, in some applications, it may be advantageous to replace the human labour and possible human error arising with analytic differentiation by the use of 'numerical differentiation'. The proposed approach might be described as an *infinitesimal jackknife* method since it perturbs the weight given to each sample observation by an infinitesimal amount to determine the approximating linear statistic. The derivative with respect to a weight in the proposed approach may be approximated numerically by a finite difference approach in which the statistic is recalculated with the weight perturbed by a finite amount for each observation in turn. This approach may be described as a *jackknife* method of linearization. A conventional approach would be to

change each weight to zero in turn, perhaps standardizing for unequal weights as in (1.15). It does not seem essential to replace the original weight by zero and, in principle, each weight might be perturbed in some other way, for example by reducing it by a fixed amount δ , smaller than the minimum value of $d_k(s)$. It seems likely that in many applications the variance estimator arising from such jackknife linearization will have very similar statistical properties to that constructed by the proposed approach. The choice between the estimators is likely to depend more on practical and computational considerations.

My final comments are on terminology. There are practical reasons why it may be helpful to give the z_k variable a name. In particular, this may be helpful for the practitioner who, for some complex statistics, has to employ two separate computational steps: (a) construction of the z_k variable, for example using least squares routines when calibration weighting is used, and (b) use of standard variance estimation software for linear statistics. Different names are used for z_k in the literature. Woodruff (1971) is usually acknowledged as the first paper in the survey sampling literature to draw attention to the role of z_k and Andersson and Nordberg (1994) refer to z_k as the *Woodruff transformation*. Woodruff and Causey (1976) refer to the approximating linear statistic as the *linear substitute* and z_k as the *substitute variable*. In the more mainstream statistical literature, Davison and Hinkley (1997, page 46) refer to the z_k as the *empirical influence values*. The term *linearized variable*, as used by Deville (1999), seems to me a simple and natural one. It is consistent with the use of the term *linearized statistic* to denote the approximating linear statistic and the term *linearization* for the method (which is a more suitable general term than Taylor series method for the broad class of approaches considered here).

ADDITIONAL REFERENCES

- DAVISON, A.C., and HINKLEY, D.V. (1997). *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.
- EFRON, B., and HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika*. 65, 457-487.
- FULLER, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*. 28, 5-23.
- WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*. 66, 411-414.
- WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*. 71, 315-321.

Response from the Authors

1. INTRODUCTION

We thank the three discussants, Phillip Kott, Babubhai Shah and Chris Skinner, for their insightful comments. Our rejoinder will attempt to address some of the issues raised by the discussants. The main aim of our paper was to study variance estimation for calibration estimators of population totals and nonlinear parameters, θ , defined as solutions to “census” estimating equations. We proposed a new Taylor linearization approach that provides a unique variance estimator, by avoiding initial evaluation of the linearized statistic at the population values. We have also shown that the variance estimator satisfies some desirable considerations, such as approximate model unbiasedness and validity under a conditional repeated sampling frame work, at least in a number of important cases. We have also shown that in two-phase sampling the variance estimator makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

Kott

Kott’s discussion focused on three applications in our paper: (i) the jackknife linearization variance estimator, v_{JL} , of the ratio estimator $\hat{Y}_R = (\bar{y}/\bar{x})X$ in simple random sampling mentioned in section 1; (ii) the general class of regression calibration weights considered in section 3.2; (iii) the general class of calibration weights studied in section 3.4. Regarding (i), we noted the result that v_{JL} is both asymptotically design unbiased and approximately model unbiased under the ratio model $E_m(y_k) = \beta x_k$ and $V_m(y_k) = \sigma^2 x_k$. Kott is correct in saying that the model bias may not be negligible if the sampling fraction, n/N , is not small. If n/N is “ignorably small”, then model unbiasedness is, in fact, valid under a general variance function $V_m(y_k) = \sigma_k^2$, as noted by Kott and previously by Särndal *et al.* (1989). Under the ratio model, Kott proposes a more appropriate variance estimator, v_m , that is model unbiased even if n/N is not small and also valid under repeated sampling. The leading terms of v_m and v_{JL} are identical, and our new approach captures only the leading term. It should be noted that model-unbiasedness of v_m depends on the validity of the assumption $\sigma_k^2 = \sigma^2 x_k$.

Turning to (ii), we have shown in section 3.2 that if the general class of regression calibration weights, (3.7), are used, our approach leads to a variance estimator that is quite complex, involving third and fourth order moments of the design weights $d_k(s)$ with $d_k(s) = 0$ if the k^{th} population element is not in the sample s . Kott proposes an attractive choice of weights obtained by replacing $c_k x_k$ in the GREG

weight (3.1) with $q_{(1)k} = \sum_{l=1}^N (\pi_{kl} - \pi_k \pi_l) x_l / (\pi_k \pi_l)$. This choice gives a variant of the “optimal” linear regression estimator and also avoids the complexities associated with the variance estimator based on the weights (3.7). This is an interesting and useful proposal, but $q_{(1)k}$ requires the knowledge of the x -vector for all the population elements, unlike (3.7) which depends only on the population total X ; in practice, only X may be available. Moreover, $q_{(1)k}$ depends on all the $N(N-1)/2$ joint inclusion probabilities π_{kl} and hence computation of $q_{(1)k}$ may become cumbersome when the sampling design is based on unequal probability sampling without replacement.

Turning to (iii), Kott proposes a generalization of the calibration weights $w_k(s) = d_k(s) F(x_k^T \hat{\lambda})$ in section 3.4 by replacing x_k with “instrumental” variables q_k having the same dimension as x_k . The corresponding z -variable in the variance estimator $v(z)$ is similar to our (3.21) with $x_k x_k^T$ and $x_k y_k$ in \hat{B}_λ changed to $q_k x_k^T$ and $q_k y_k$ respectively and $F(x_k^T \hat{\lambda})$ changed to $F(q_k^T \hat{\lambda})$. This is an useful extension. Kott notes that \hat{B}_λ remains a model unbiased estimation of B_λ if $f(q_k^T \lambda)$ in \hat{B}_λ is replaced by any constant and the resulting z_k is unaffected asymptotically under repeated sampling. However, Kott also notes that the term $f(q_k^T \lambda)$ can matter even asymptotically if the calibration is used to adjust for unit nonresponse by treating sample response as a second phase of sampling. Using the result for two-phase sampling given in the Appendix, Kott then obtains a corresponding variance estimator, $v(\hat{Y}_{GC})$. This extension for nonresponse setting is also useful. It is indeed surprising that the second term in $v(\hat{Y}_{GC})$ provides the model based correction he recommended for the ratio estimator \hat{Y}_R under simple random sampling in the absence of nonresponse.

Finally, Kott raises a question on the customary “plug-in” or “substitution” method used for variance estimation, as done in (2.4), where we plug in z_k for \tilde{z}_k . He asks if it is legitimate to plug in z_k for \tilde{z}_k , where $z_k - \tilde{z}_k = O_p(1/\sqrt{n})$, when they are $n(n-1)/2$ terms in the variance estimator $v(\tilde{z}_k)$, as in the case of Sen-Yates-Grundy variance estimator. We are not sure if we have understood his point correctly, but as long as $O_p(1/\sqrt{n})$ is uniform in k , say a/\sqrt{n} , then $v(z) = v(\tilde{z}) + \text{lower order terms}$.

Shah

Shah’s prescription (steps 1-4) clearly summarizes our method. Shah also notes that his steps 1 and 2, leading to our z -variable, produces the “correct” linearization in many other important applications not studied in our paper,

including Wilcoxon two sample rank sum test and estimation of regression coefficients and hazard rate in the Cox proportional hazard model. Shah's unpublished paper (seen by courtesy of the author) spells out the z -variable for those applications, but using design weights. Extension to calibration weights should follow along the lines of section 3.

Shah makes an important point that step 3 for the computation of the variance estimate is independent of the linearization in step 1 and 2 and that it is "aptly demonstrated by the discussion on two-phase sampling in section 4". He also notes that for two-phase sampling, linearization (step 1) can be performed using only the first-phase H-T weights π_{1k}^{-1} , by treating the second phase weights, $\pi_{2k/1}^{-1}$, if $k \in s$ and 0 if k is not in the second-phase sample s as data, and that the resulting step 2 produces the same approximation as given in our paper. We have verified this equivalence result for the two-phase ratio estimator in Example 4.1, and it is likely to hold generally. Shah's proposal might simplify the implementation of step 1 to some extent.

Skinner

Skinner gives a clear appraisal of our linearization method and raises a number of important points: (i) terminology, (ii) possible extensions to non-smooth statistics such as quantiles, (iii) modifications of the method to handle unit nonresponse, (iv) possible use of numerical differentiation to calculate the z_k -variables.

With regard to point (i), Skinner notes that it would be useful to give the z_k variable a name since different names have been used in the literature. He suggests that the term *linearized variable*, as used by Deville (1999), is a simple and natural one since it is consistent with the usage of *linearized statistic* to denote the approximating linear statistic and linearization for the method. We are in agreement with Skinner's suggestion.

Turning to point (ii), a difficulty in extending our proposal to nonsmooth statistics $\hat{\theta} = f(d(s))$, such as quantiles, is that $f(\cdot)$ is not a differentiable function. A way to get around this difficulty is to approximate $\hat{\theta} - \theta$ by a differentiable function and then apply our method to the approximation. For example, in the case of the p^{th} quantile θ , Francisco and Fuller (1991) and Shao (1991) established the following asymptotic approximation valid for stratified multistage designs:

$$\hat{\theta} - \theta \approx -\frac{1}{h(\theta)} \left\{ \hat{F}_w(\theta) - p \right\},$$

where $\hat{F}_w(\theta) = \sum w_k(s) I(y_k \leq \theta) / \sum w_k(s)$ is the calibration estimator of the distribution function $F(\cdot)$ at θ , $F(\theta) = N^{-1} \sum I(y_k \leq \theta) = p$, and $h(\theta)$ is the value of the density

function $h(\cdot)$ at θ . The definition of $h(\cdot)$ requires reference to a sequence of populations (Shao and Rao 1993) or to a superpopulation (Francisco and Fuller 1991). We used $h(\cdot)$ to denote the density rather than the customary $f(\cdot)$ because we used $f(d(s))$ to denote the estimator $\hat{\theta}$. Now, suppose $w_k(s) = d_k(s) g_k(d(s))$, where $g_k(d(s))$ is the GREG weight given by (3.1). We can then use (3.2) and (3.3) to get the linearized variable z_k from the above approximation to $\hat{\theta} - \theta$, by replacing $h(\theta)$ with a suitable estimator $\hat{h}(\hat{\theta})$; for example the kernel-based estimator of $h(\cdot)$ used by Berger and Skinner (2003). Similarly, one can apply the method to general calibration weights, $w_k(s)$, using the results of section 4. Variance estimators of a low income proportion, say $\theta = F(\tau/2)$ where τ is the median income, can also be obtained using the asymptotic approximation for $\hat{\theta} - \theta$ developed by Shao and Rao (1993). Berger and Skinner (2003) studied variance estimation for a low income proportion when generalized raking ratio weights, $w_k(s)$, are used. We can apply the results in section 3.2 to this case, and the resulting linearized variable z_k will account for the calibration. Also, it will be different from the Deville z -variable (10) in Berger and Skinner (2003).

The modification suggested in point (iii) to handle unit nonresponse is very important, and it broadens the applicability of our method. As noted by Skinner, Kott and Fuller (2002), it is important to retain the g -weights in variance estimation whenever the limiting values of the estimators \hat{X} differ from the corresponding control totals X , as in the case of non-response or non-coverage. Our method automatically accounts for the g -weights and may lead to consistent variance estimators in such cases. Empirical results of Skinner with d'Arrigo in this context are very interesting. The case of variance estimators for alternative calibration estimators, studied in section 3.4, relative to customary variance estimators that replace $d_k(s) f(x_k^T \hat{\lambda})$ in the expression for \hat{B}_k by $d_k(s)$ or $w_k(s)$ need further study, as noted by Skinner.

It may be noted that unit nonresponse is typically treated as second phase sampling (e.g., Poisson sampling with unknown response probabilities) and Skinner notes that our method may lead to consistent variance estimators even when the estimators are based only on the sampling inclusion probabilities. However, control totals X are needed to get valid estimators of the total Y , under some assumptions on the response probabilities (Fuller 2002, equation (8.4)). We have extended our method to handle weight adjustment for unit nonresponse and imputation for item nonresponse when control totals are not available, assuming uniform response within classes (Demnati and Rao 2002). The resulting variance estimators are naturally more complex compared to Skinner's modification for unit nonresponse in the presence of control totals.

Turning to point (iv) on the possible use of numerical differentiation to calculate the linearized variables z_k , Woodroff and Causey (1976) used such a method to calculate the derivatives $\partial g(\mathbf{a})/\partial a_i|_{\mathbf{a}=\hat{\mathbf{y}}}$ given in (1.4) when $\hat{\theta} = g(\hat{\mathbf{Y}})$. Skinner proposes perturbing each weight $d_k(s)$ in turn and then recalculating $\hat{\theta}$; for example, by replacing it by a fixed amount δ smaller than the minimum value of $d_k(s)$, $k \in s$. He conjectures that the proposed approach should lead to variance estimators very similar to those obtained through analytical differentiation. It would be useful to study the statistical properties of the proposed approach to analytic differentiation of $f(\mathbf{d}(s))$ with respect to weights $d_k(s)$.

We hope the discussions by Kott, Shah and Skinner will stimulate further work on the approach to variance estimation presented in our paper.

REFERENCES

- BERGER, Y.G., and SKINNER, C.J. (2003). Variance estimation for a low income proportion. *Applied Statistics*, 52, 457-468.
- DEMNATI, A., and RAO, J.N.K. (2002). Linearization variance estimators for survey data with missing responses. *Proceeding of the Section Survey Research Methods*, American Statistical Association, 736-740.
- FRANCISCO, C.A., and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.
- SHAO, J. (1991). L-statistics in complex problems. Technical Report, University of Ottawa, Ottawa.
- SHAO, J., and RAO, J.N.K. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhyā*, Series B, 55, 393-414.
- WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

Weighting Sample Data Subject to Independent Controls

CARY T. ISAKI, JULIE H. TSAY and WAYNE A. FULLER¹

ABSTRACT

In the U.S. Census of Population and Housing, a sample of about one-in-six of the households receives a longer version of the census questionnaire called the long form. All others receive a version called the short form. Raking, using selected control totals from the short form, has been used to create two sets of weights for long form estimation; one for individuals and one for households. We describe a weight construction method based on quadratic programming that produces household weights such that the weighted sum for individual characteristics and for household characteristics agree closely with selected short form totals. The method is broadly applicable to situations where weights are to be constructed to meet both size bounds and sum-to-control restrictions. Application to the situation where the controls are estimates with an estimated covariance matrix is described.

KEY WORDS: Raking; Regression; Quadratic programming; Coverage adjustment; Integer weights; Weighting area.

1. INTRODUCTION

Given the availability of known characteristic totals, it is common among survey practitioners to use such information in estimators of the post stratified, ratio and regression type. The known characteristic totals are sometimes called independent controls because they are derived outside of the survey situation. Use of independent controls tends to reduce the variance of most estimates. Independent controls also often compensate for coverage problems in surveys. See Deville and Särndal (1992) and Fuller (2002).

The U.S. decennial census utilizes a sample for the measurement of selected characteristics. The questionnaire for these characteristics is called the long form and the sample for the long form consists of a random sample of addresses. The long form questionnaire requests information that is asked of all individuals (called short form information) plus information on a set of additional characteristics. In previous Censuses, raking to controls based on short form information was used to construct weights for the long form sample. Two sets of sample weights were created, one for person characteristics and one for housing unit characteristics.

The set of categories used for person weighting was a classification of individuals by race, Hispanic origin, age and sex, family type, and household size. For households, the categories were the cross classification of race by Hispanic-origin-of-householder by tenure by household type and size. In the 1990 Census long form weighting process, persons and housing units were each classified by four sets of classifications for raking in four dimensions. When raking was completed, the long form sample weights were converted to integers. Integer weights are desirable because,

unlike real weights, integer weights provide arithmetically consistent totals of integral characteristics. For details, see Schindler, Griffin and Swan (1992).

Long form weighting using short form census information is a part of the Canadian Census of population and housing. Unlike the procedure used by the U.S. Census Bureau (USCB), the procedure used at Statistics Canada constructs a single set of household weights using regression estimation. See Bankier, Houle and Luc (1997). Should the initial weights generated by the regression procedure exceed prescribed bounds, collapsing of cells defining explanatory variables is carried out. Linear dependencies and near linear dependencies among the explanatory variables are also removed by eliminating variables. See Bankier, Rathwell and Majkowski (1992).

Lemaître and Dufour (1987) used a generalized least squares estimator (GLS) to construct weights meeting person and household constraints. Alexander (1987) considers a procedure for constructing household weights in the census setting. One of his distance functions is similar to the one used in this paper.

The use of quadratic programming to compute regression weights in the survey context was suggested by Husain (1969). An application of quadratic programming (QP) in a Census environment is that in Isaki, Ikeda, Tsay and Fuller (2000) where household weights for Census households were obtained using person totals as controls. Motivation for the use of various distance functions can be found in these two papers and in Deville and Särndal (1992) who discuss a general class of estimators called calibration estimators. Fuller, Laughin and Baker (1994) consider a regression weight generation procedure that is modified so that all weights are positive and very large weights are made

¹ Cary T. Isaki and Julie H. Tsay, U.S. Bureau of the Census, Statistical Research Division, Washington, D.C. 20233, U.S.A. E-mail: Julie.Hsu.Ling.L.Tsay@census.gov; Wayne A. Fuller, Iowa State University, Department of Statistics, 221 Snedecor Hall, Ames, Iowa 50011, U.S.A.

smaller than the corresponding least squares weight. Jayasuriya and Valliant (1996) also consider a restricted regression. Fuller (2002) is a review of regression estimation.

Our proposed long form weighting method is a type of regression estimation and, like the Statistics Canada approach, provides a single set of household weights that maintain given independent controls. We generate household weights using quadratic programming with the restrictions that the weights fall within a specified range and that the weights maintain control totals. In the following, we refer to the suggested method as the quadratic programming method or QP.

2. THE QUADRATIC PROGRAMMING METHOD

The purpose of quadratic programming is to produce sample weights that i) are close to initial weights, ii) are within reasonable bounds, iii) maintain specified control totals and iv) produce a design consistent estimator. Apart from the bounds on the weight, the weights from quadratic programming are those of a simple regression estimator. We first describe the mathematical form of the QP and then discuss the implementation. Let

- i) $\{W_i; i = 1, 2, \dots, n\}$ denote the set of final housing unit weights, where i denotes the i^{th} long form sample household and n is the size of the long form sample,
- ii) $\{W_i^{(2)}; i = 1, 2, \dots, n\}$ denote the set of initial housing unit weights,
- iii) $X_{ji}, j = 1, 2, \dots, m_p, i = 1, 2, \dots, n$; denote the observation on the j^{th} person control variable for the i^{th} sample household,
- iv) $Z_{ji}, j = 1, 2, \dots, m_h, i = 1, 2, \dots, n$; denote the observation on the j^{th} household control variable for the i^{th} sample household,
- v) $X_j, j = 1, 2, \dots, m_p$, denote the j^{th} person control,
- vi) $Z_j, j = 1, 2, \dots, m_h$, denote the j^{th} household control.

The quadratic programming method seeks $W_i, i = 1, 2, \dots, n$, that minimize a quadratic objective function subject to linear constraints. In our application we minimize

$$g(W) = \sum_{i=1}^n (W_i - W_i^{(2)})^2 [W_i^{(2)}]^{-1}, \quad (1)$$

subject to

$$\sum_{i=1}^n W_i X_{ji} = X_j, \quad \text{for } j = 1, 2, \dots, m_p, \quad (2)$$

$$\sum_{i=1}^n W_i Z_{ji} = Z_j, \quad \text{for } j = 1, 2, \dots, m_h, \quad (3)$$

$$1 \leq W_i \leq K \quad (4)$$

where the summations are over housing units in the long form sample. Observe that the long form household weights are bounded below by one. This is on the basis that an element in the sample should at least "represent" itself. In our program, K was set equal to 48 but the bound was never attained. The lower bound of one was attained. The FORTRAN subroutine from IMSL was used to solve the QP. Other programs, such as LCP of SAS®/IML, are available.

The USCB's current long form weighting procedure rakes the initially weighted long form sample counts to the census counts for the control categories. The weighting is done by subdivisions of the country called weighting areas and is done separately for person and household characteristics. The nominal sample rates for the long form are one-in-two, one-in-six, and one-in-eight. The nominal sampling weights are the inverses of the nominal sampling rates and are denoted by $W_i^{(1)}$. A second set of weights, denoted by $W_i^{(2)}$, are the realized sampling rates calculated for cells, where the cells are required to contain at least five sample households. For details on the USCB's procedures see Schindler *et al.* (1992).

Since we intend to compare the raking and QP methods, we use most of the USCB's person and household categories as the X_j and Z_j control totals in the quadratic program, but some changes were instituted. For example, while we maintained all of the age-race-sex person categories, we did not use a category based on the nominal sampling rates.

We used the USCB's specifications for determining whether a cell category would be retained as a separate control or would be combined with another cell and we used the USCB's procedure for determining the cells to be combined. This capitalized on the USCB's experience and minimized differences between the USCB's set of long form control totals and the set used by the QP method. The procedure used to define $W_i^{(2)}$ is given in the appendix.

Two possibilities exist for the control totals to be used in the construction of weights for the long form of the U.S. 2000 Census. One possibility is to use controls from the 2000 Census short form. That is, the independent controls to be maintained in long form weighting are those that are tabulated from the Census short form. When the Census is used as the control, the person control (X_j) categories include a cross classification of age and sex-race/ethnicity. Other characteristics, such as tenure, were used as additional

controls. The majority of the household control categories (Z_j) are defined by a cross classification of household type (e.g., family with children under 18) and household size (e.g., number of persons in the family). The Z_j also include race/ethnicity of the householder cross-classified by tenure.

The other possible set of controls for the 2000 Census is the set of estimates from the post enumeration survey, called the Accuracy and Coverage Evaluation (A.C.E.) survey. The A.C.E. survey is designed to estimate person characteristics only. The X_i for the A.C.E. include age-sex-race/ethnicity-tenure controls.

The last step in long form weighting is to round the W_i to integers. Integer weights prevent discrepancies between sets of estimates caused by rounding of real valued estimates. Sample housing units were grouped by race/ethnicity of the householder and by tenure. Then within each group, the sample was sorted by family type by household size. The weights were then rounded to integers using the cumulate-and-round procedure. Table 1 illustrates the method. The partial sums of the weights are formed (cumulated) as shown in the column CW. The partial sums are then rounded as shown in the column RCW. The integer weight for element i is the difference between successive entries $i - 1$ and i in the RCW column.

Table 1
Illustration of Cumulate and Round

Sample Unit	Initial Weight	CW	RCW	Integer Weight
1	3.333	3.333	3	3
2	2.500	5.833	6	3
3	1.428	7.261	7	1
4	1.250	8.511	9	2
5	1.111	9.622	10	1
6	5.021	14.643	15	5

3. VARIANCE ESTIMATION

Variances of long form estimates were estimated using the jackknife method. In the numerical results using census controls, sixteen replicates were formed. Sixteen was chosen for convenience and a larger number could have been used. The long form sample was ordered by the census identification number within blocks and sixteen replicates were formed as the sixteen one-in-sixteen systematic samples. Sixty seven replicates were formed for the estimates using ACE controls.

3.1 Replicates for Census Controls

The jackknife replicate is created by deleting the i^{th} group of elements, computing the quadratic programming weights and rounding the weights to integers. Because of the rounding, the usual jackknife variance estimation procedure required modification. To isolate the effect of rounding, we consider the replicate estimate constructed with real-valued weights. Let

- $\hat{\theta}_w$ = the sample estimator with weights rounded to integers,
- $\hat{\theta}_R$ = the sample estimator with real-valued weights,
- $\hat{\theta}_{R(i)}$ = jackknife replicate estimate with i^{th} group deleted and real-valued weights,
- $\hat{\theta}_{w(i)}$ = jackknife replicate estimate with i^{th} group deleted weights rounded to integers,

and let

$$\bar{\theta}_w = r^{-1} \sum_{i=1}^r \hat{\theta}_{w(i)}, \tag{5}$$

where r is the total number of replicates. Then the jackknife deviation for the estimator with integer weights can be decomposed as

$$\begin{aligned} \hat{\theta}_{w(i)} - \bar{\theta}_w &= \hat{\theta}_{R(i)} - \hat{\theta}_R \\ &\quad + [\hat{\theta}_{w(i)} - \bar{\theta}_w - (\hat{\theta}_{R(i)} - \hat{\theta}_R)]. \end{aligned} \tag{6}$$

We assume that the error in the rounding operation is independent of the group chosen for deletion, a reasonable assumption, given that the deletion produces an entire new set of weights to be rounded. Then

$$\begin{aligned} E\{(\hat{\theta}_{w(i)} - \bar{\theta}_w)^2\} &\doteq E\{(\hat{\theta}_{R(i)} - \hat{\theta}_R)^2\} \\ &\quad + E\{[(\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}) - (\bar{\theta}_w - \hat{\theta}_R)]^2\}. \end{aligned} \tag{7}$$

Assume that the average of the $\hat{\theta}_{R(i)}$ is equal to $\hat{\theta}_R$. Then the last term of (7) is a replicate deviation for the difference between the real and rounded estimates. Then

$$\begin{aligned} E\{[(\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}) - (\bar{\theta}_w - \hat{\theta}_R)]^2\} &= \\ r^{-1}(r-1)V\{\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\} &= V\{\hat{\theta}_w - \hat{\theta}_R\} \end{aligned} \tag{8}$$

where $V\{\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\}$ is the variance due to rounding for a sample of $r - 1$ groups and $V\{\hat{\theta}_w - \hat{\theta}_R\}$ is the variance due to rounding for a sample of r groups. In obtaining (8) we assumed the variance due to rounding for a sample of r groups is the variance for $r - 1$ groups multiplied by $r^{-1}(r - 1)$. Thus

$$E \left\{ (r-1)^{-1} \sum_{i=1}^r (\hat{\theta}_{w(i)} - \bar{\theta}_w)^2 \right\} \doteq E \left\{ (r-1)^{-2} r \hat{V}_R \{ \hat{\theta}_R \} + V \{ \hat{\theta}_w - \hat{\theta}_R \} \right\}, \quad (9)$$

where

$$\hat{V}_R \{ \hat{\theta}_R \} = r^{-1} (r-1) \sum_{i=1}^r (\hat{\theta}_{R(i)} - \hat{\theta}_R)^2$$

is the jackknife variance estimator for the estimator with real weights. Then an estimator of the variance due to rounding is

$$\begin{aligned} \hat{V} \{ \hat{\theta}_w - \hat{\theta}_R \} &= r^{-1} (r-1) \left[(r-1)^{-1} \sum_{i=1}^r (\hat{\theta}_{w(i)} - \bar{\theta}_w)^2 - r (r-1)^{-2} \hat{V}_R \{ \hat{\theta}_R \} \right] \\ &= r^{-1} \left[\sum_{i=1}^r (\hat{\theta}_{w(i)} - \bar{\theta}_w)^2 - r (r-1)^{-1} \hat{V}_R \{ \hat{\theta}_R \} \right]. \end{aligned} \quad (10)$$

Based on these results, the estimated variance for the rounded estimator is

$$\begin{aligned} \hat{V} \{ \hat{\theta}_w \} &= (r-1)^{-1} (r-2) \hat{V}_R \{ \hat{\theta}_R \} \\ &\quad + r^{-1} \sum_{i=1}^r (\hat{\theta}_{w(i)} - \bar{\theta}_w)^2. \end{aligned} \quad (11)$$

3.2 Replicates for A.C.E. Controls

The replicates for estimates constructed with A.C.E. controls were modified so that the estimated variances contained a component for the error in the A.C.E. estimates. The data in a weighting area were assigned to 67 replicates where 67 is the number of controls. The procedure requires the number of replicates to equal or exceed the number of controls if the covariance matrix of the estimated control totals is to be reproduced. More replicates than controls can be used. See Fuller (1998).

The estimator of the total of a characteristic for the long form is a type of regression estimator using the A.C.E. numbers as controls. We write the estimator for the total based on real valued weights as

$$\hat{\theta}_R = \hat{\mathbf{X}}_A \hat{\boldsymbol{\beta}}, \quad (12)$$

where $\hat{\mathbf{X}}_A$ is the vector of A.C.E. estimates and $\hat{\boldsymbol{\beta}}$ is the regression coefficient computed with the long form data.

Let $\hat{\mathbf{V}}_{AA}$ be the $r \times r$ covariance matrix of the vector of A.C.E. controls, where $\hat{\mathbf{V}}_{AA}$ is estimated as part of the A.C.E. process, and $r = 67$. Let $\lambda_1, \lambda_2, \dots, \lambda_r$ be the roots of $\hat{\mathbf{V}}_{AA}$ and let

$$\mathbf{Q}' \mathbf{V}_{AA} \mathbf{Q} = \mathbf{\Lambda}, \quad (13)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$, and \mathbf{Q} is the matrix composed of the characteristic vectors of $\hat{\mathbf{V}}_{AA}$. Recall that

$$\hat{\mathbf{V}}_{AA} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}'$$

and

$$\hat{\mathbf{V}}_{AA} = \sum_{j=1}^r \mathbf{q}_{\bullet j} \lambda_j \mathbf{q}'_{\bullet j} = \sum_{j=1}^r \mathbf{z}_{\bullet j} \mathbf{z}'_{\bullet j} \quad (14)$$

where $\mathbf{q}_{\bullet j}$ is the j^{th} column of \mathbf{Q} and $\mathbf{z}_{\bullet j} = \lambda_j^{1/2} \mathbf{q}_{\bullet j}$.

Using result (14), controls for the r replicates were constructed as

$$\ddot{\mathbf{X}}_{A(i)} = \hat{\mathbf{X}}_A + c \mathbf{z}'_{\bullet i}, \quad i = 1, 2, \dots, r, \quad (15)$$

where $\hat{\mathbf{X}}_A$ is the row vector of the original controls and c is a constant. The constant c is determined so that the expectation of the sum of the jackknife squared deviations for the elements of the vector \mathbf{X} are the diagonal elements of $\hat{\mathbf{V}}_{AA}$. In our application, the constant c is $(r-1)^{-1/2} r^{1/2}$ and

$$\begin{aligned} (r-1) r^{-1} \sum_{j=1}^r c^2 \mathbf{z}_{\bullet j} \mathbf{z}'_{\bullet j} &= \sum_{j=1}^r \mathbf{z}_{\bullet j} \mathbf{z}'_{\bullet j} = \hat{\mathbf{V}}_{AA}. \end{aligned} \quad (16)$$

Thus, if the characteristic being “estimated” is one of the controls used in the QP, the jackknife procedure returns the A.C.E. estimated variance for that characteristic. The $\mathbf{z}_{\bullet j}$ are assigned at random to the r replicates.

Using the regression representation, we write the estimator for the i^{th} replicate as

$$\begin{aligned} \ddot{\theta}_{R(i)} &= \ddot{\mathbf{X}}_{A(i)} \hat{\boldsymbol{\beta}}_{(i)} \\ &= \hat{\mathbf{X}}_A \hat{\boldsymbol{\beta}}_{(i)} + (\ddot{\mathbf{X}}_{A(i)} - \hat{\mathbf{X}}_A) \hat{\boldsymbol{\beta}}_{(i)} \\ &= \hat{\theta}_{R(i)} + c \mathbf{z}'_{\bullet i} \hat{\boldsymbol{\beta}}_{(i)}, \end{aligned} \quad (17)$$

where $\ddot{\theta}_{R(i)}$ is the real-valued estimator computed with the i^{th} group deleted using $\ddot{\mathbf{X}}_{A(i)}$ as the control vector, $\hat{\boldsymbol{\beta}}_{(i)}$ is the regression coefficient computed with the i^{th} group deleted, and $\hat{\theta}_{R(i)}$ is the real-valued estimator computed with the i^{th} group deleted using $\hat{\mathbf{X}}_A$ as the control vector. Then

$$\ddot{\theta}_{R(i)} - \hat{\theta}_R = \hat{\theta}_{R(i)} - \hat{\theta}_R + c \mathbf{z}'_{\bullet i} \hat{\boldsymbol{\beta}}_{(i)}.$$

Because $\mathbf{q}_{\bullet j}$ are assigned to replicates at random, the expectation of the replicate variance estimator for the real-valued estimator based on A.C.E. controls is

$$\begin{aligned}
E\{\hat{V}_R(\hat{\theta}_R)\} &= E\left\{r^{-1}(r-1)\sum_{i=1}^r(\ddot{\theta}_{R(i)} - \hat{\theta}_R)^2\right\} \\
&= E\left\{r^{-1}(r-1)\sum_{i=1}^r(\hat{\theta}_{R(i)} - \hat{\theta}_R)^2\right\} \\
&\quad + E\{\hat{\beta}'_{(i)}\hat{V}_{AA}\hat{\beta}_{(i)}\}. \quad (18)
\end{aligned}$$

Now, assuming $E\{\hat{V}_{AA}\} = \mathbf{V}_{AA}$, $E\{\hat{\beta}_{(i)}\} = \beta$, and that \hat{V}_{AA} is independent of $\hat{\beta}_{(i)}$,

$$\begin{aligned}
E\{\hat{\beta}'_{(i)}\hat{V}_{AA}\hat{\beta}_{(i)}\} &= \beta'\mathbf{V}_{AA}\beta \\
&\quad + \text{tr}\{V\{\hat{\beta}_{(i)}\}\mathbf{V}_{AA}\},
\end{aligned}$$

where $\text{tr}\{\mathbf{V}_{AA}\}$ is the trace of the matrix. It follows that

$$\begin{aligned}
E\left\{r^{-1}(r-1)\sum_{i=1}^r(\ddot{\theta}_{R(i)} - \hat{\theta}_R)^2\right\} \\
= E\left\{r^{-1}(r-1)\sum_{i=1}^r(\hat{\theta}_{R(i)} - \hat{\theta}_R)^2\right\} \\
+ \beta'\mathbf{V}_{AA}\beta + O(n^{-2}), \quad (19)
\end{aligned}$$

where we assume $\text{tr}\{\mathbf{V}_{AA}\} = O(n^{-1})$ and $\text{tr}\{V\{\hat{\beta}_{(i)}\}\} = O(n^{-1})$, where n is the sample size. The first term on the right of the equality in (19) is the expectation of the variance estimator for the variance due to the sampling of long forms from the census. The second term is the contribution of the variance of the error in the A.C.E. estimates to the total variance. Thus, the variance estimator based on $\ddot{\theta}_{R(i)}$ estimates both components of variation. Observe that the estimated covariance matrix for the controls is \hat{V}_{AA} , as it should be.

4. NUMERICAL RESULTS

We used the USCB's 1990 Census data file to illustrate the application of the QP method to actual data. The file provides data for households and for persons in households, together with long form weights as developed for the 1990 U.S. Census. Hence, the file provides data appropriate for comparing the performance of the USCB's 1990 long form weighting method with the QP method.

The USCB long form sample weighting is done by weighting area, where the weighting areas usually contain two to three thousand housing units. There were about 56,000 weighting areas in the U.S. in 1990. For our numerical work we chose weighting area (WA) 1788 that contains 8,034 occupied housing units and 25,145 persons.

In Table 2 we provide estimates of some person and housing unit characteristics for weighting area 1788. The

characteristics in the table, except the number of rented units, were suggested by subject matter personnel at the USCB. In Table 2, Est.(H) is the long form sample weighted estimate computed with housing unit weights, Est.(P) is the long form sample weighted estimate computed with person weights. The quadratic programming estimator constructed with Census controls is called QP in the table, while QPG is used to denote the generalization of the quadratic programming estimator with objective function (20). The QPG estimator is discussed subsequently. The USCB housing unit estimates in Table 2 that are based on person weights were created by using the householder weight as the housing unit weight. Every occupied unit contains a single householder. The householder procedure is called the *principal person method* by Alexander (1987). All estimates in the table are given as a percent of the census count.

Estimates constructed by the two USCB methods can differ by several percentage points with the differences between Est.(P) and Est.(H) for rented units, persons aged 0 to 4 years, persons aged 65 and over, Hispanic, Asian, and persons in rented units being noticeable. The Est.(H) estimate for persons in rented units is closer to 100 than the Est.(P) estimate.

The cell collapsing rules produced 45 person and 22 housing unit controls for WA 1788. An example of a person control is the total number of Non-Hispanic Black males aged 65 and over, while an example of a housing unit control is the total number of Non-Hispanic White owned housing units. Total Black persons is an implicit control in WA 1788. Controls for total persons 18-44, total persons 45-64, total males, total renters and total number of rented housing units were added to the QP. Apart from the controls mentioned above, none of the remaining characteristics in Table 2 is also used as a control in the QP procedure.

The QP estimates and standard errors of the QP estimates are given, as a percent of the census counts, in the fourth and fifth columns of Table 2. The agreement between count and QP estimates for household characteristics are comparable to the USCB household based estimates and superior to USCB person based estimates. For person counts, the QP estimates are generally closer to the census counts than either of the USCB raking estimates.

The largest difference between a QP estimate and the census count relative to the standard error is for the estimate of the number of households with own children present, where the difference is about 1.6 standard errors. The majority of the QP estimates differ from the census count by less than one standard error. A number of the USCB person estimates deviate from the census count by more than one QP standard error.

Table 2
Estimated Occupied Housing Unit and Person Characteristics for WA 1788

	Census Count	Est.(H)* Count (%)	Est.(P)** Count (%)	QP† Count (%)	se (QP) Count (%)	QPG†† Count (%)	se (QPG) Count (%)
Housing unit characteristics							
With Own Children	4,349	100.18	100.45	100.21	0.13	100.18	0.14
Not With Own Children	3,685	99.78	99.67	99.76	0.15	99.78	0.16
With 1 to 4 Persons	6,785	100.00	100.57	100.04	0.05	100.07	0.05
With 5+ Persons	1,249	100.00	97.51	99.76	0.30	99.60	0.30
Rented Unit	2,559	100.00	95.97	100.00	0.19	99.92	0.16
Owned Unit	5,475	100.00	102.02	100.00	0.09	100.04	0.08
Person characteristics							
Age 0-4 years	2,493	101.92	97.95	98.84	1.68	99.96	0.29
Age 5-17 years	6,339	103.91	101.07	100.63	0.71	99.98	0.18
Age 18-44 years	12,711	99.50	99.69	100.01	0.05	100.00	0.06
Age 45-64 years	3,028	101.65	101.95	99.90	0.09	99.97	0.09
Age 65+ years	574	81.18	93.73	100.17	0.85	100.00	0.27
Males	12,473	99.95	99.64	100.06	0.08	99.98	0.09
Females	12,672	101.43	100.36	99.95	0.10	100.01	0.09
Hispanic	2,385	95.38	103.40	99.96	0.38	99.87	0.38
Not Hispanic	22,760	101.25	99.64	100.03	0.07	100.00	0.10
Black	1,285	101.08	101.79	100.86	1.22	99.77	0.54
White	22,372	100.69	99.91	100.03	0.07	100.00	0.10
Asian	257	92.60	80.05	96.83	2.32	99.76	0.50
Remainder	1,231	101.94	103.89	105.84	9.54	100.78	1.75
In Rented Unit	7,978	102.04	95.41	100.01	0.24	99.92	0.19
In Owned Unit	17,167	100.06	102.13	100.00	0.09	100.02	0.13

* USCB weights for households

** USCB weights for persons

† QP weights with 82 constraints

†† Generalized QP with 13 constraints and objective function (20)

Because the number of rented units, persons aged 18-44, persons aged 45-64, males, and persons in rented units were used as controls in the QP procedure, differences between QP estimates and census totals for those categories are due to rounding. The standard errors demonstrate that the rounding can lead to sizeable deviations from the controls.

The 45 person and 22 housing unit control totals obtained by the collapsing rules are such that a margin estimate, such as total males, may not be constrained to agree with the count. In addition, for different weighting areas, USCB's collapsing procedure gives different person and housing unit constraints. Thus we considered adding some margin totals

to the set of control totals. To reduce the impact of the added controls on the weights, we replaced the original constraints with additional terms in the objective function. The terms are deviations between the final estimates and the control totals. The objective function becomes

$$G(W) = g(W) + \sum_{j=1}^{67} \alpha_j \left(\sum_i W_i X_{ji} - X_j \right)^2, \quad (20)$$

where $g(W)$ is defined in expression (1), the $\{X_{ji}, j = 1, 2, \dots, 67\}$ is the set of auxiliary variables defining the 45 person and 22 housing unit controls, and α_j are constants to be specified. The X_{ji} for category j of household i for a person characteristic is the number of individuals in category j in the housing unit. The X_{ji} for a housing unit characteristic is one if the housing unit has the characteristic and zero otherwise. In our application, the function is minimized subject to two household controls and eleven person controls. The housing unit controls are rented housing units and owned housing units. The person controls are persons 0 to 4 years, persons 5 to 17 years, persons 18 to 44 years, persons 45-64 years, persons 65 years and over, males, black, white, Asian, Hispanic, and renters. The α_j are $10[\bar{W}^{(2)}]^{-1}[\sigma_j^2]^{-1}$, where $\bar{W}^{(2)} = 8.95$ is the mean of the $W_i^{(2)}$, $\sigma_j^2 = P_j(1 - P_j)$, and P_j is the proportion of the population in cell j . The α_j would minimize the mean square error of an estimated total if there was a single control variable and the squared correlation between the control variable and the dependent variable was about 0.9. Thus, the function exerts considerable pressure for the final estimate to be close to the control total.

The QP solution to (20) gives a type of regression estimator. See Fuller (2002) and Fuller and Isaki (2001). Rao and Singh (1997) and Bardsley and Chambers (1984) consider related estimators.

Using $G(W)$ of (20) and the 13 linear constraints, the results in the final two columns of Table 2, under the heading "QPG", were obtained. As expected, the estimates are close to Census totals because the Census marginals were used as constraints. The relative percent differences between the QP estimate and the census count for the 67 characteristics in $G(W)$ of (20) ranged from -3.50% to 3.75% with about 50 of the differences being less than one percent.

The sample weights obtained by the two programming approaches are compared to those of the USCB's household raking method in Table 3. The number and type of controls used under the USCB raking was not determined exactly because the number depends on the execution of the USCB collapsing procedure and on some preliminary files that are not readily available. However, we believe the number to be about 67 because the collapsing procedure used to form the 67 cells is basically that used by the USCB. The QP

procedure used 82 controls and the QPG procedure used 90 controls. The range of weights for the two QP methods are similar with a smaller range for raking. There are modest differences among the three sums of squares of the weights. The $g(W)$ values are also similar, with the value for (20) being the largest. The $g(W)$ value is the quantity being minimized by the weights of the first line of the table. The sum of squares of the weights for the QP of (20) could be reduced by reducing the α_j in the objective function.

We also used data from the 1990 Census to simulate the situation in which the controls come from adjusted census counts. For 1990, person estimates from the 1990 Post Enumeration Survey are available, but there are no housing unit estimates based on that survey. We call these estimates A.C.E. estimates. See Hogan (1993) and Isaki, Tsay and Fuller (2000). Estimates for WA 1788 were created by the QP method, using the A.C.E. estimates as controls. We used $G(W)$ of (20) as the objective function with 63 age-race-sex-tenure person characteristics in the second term of the objective function and 11 person constraints. The person constraints are persons 0 to 4 years, 5 to 17 years, 18 to 44 years, 45 to 64 years, 65 and over, total males, total Hispanic, total Black, total White, total Asian and total persons in rented units.

Table 3
Properties of Long Form Housing Unit Sample Weights
in WA 1788

Method	Minimum Weight	Maximum Weight	$\sum_i W_i^2$	$g(W)$
QP with $g(W)$ of (1) 72 constraints	1	26.5	78,028	326
QP with $G(W)$ of (20) 13 exact constraints	1	29.9	78,672	383
Raking	4	22	77,000	369

Table 4 contains the estimates for WA 1788 identified as QPG and given as a percent of the census counts. The QPG estimates for these eleven person characteristics agree with the A.C.E. estimates, except for rounding error. The standard errors reflect the error in the A.C.E. estimates and, hence, are much larger than the standard deviation of rounding error. For example, the rounding error standard deviation for persons 18 - 44 is 0.06 in Table 2, while the standard error for the ACE estimate of persons 18 - 44 is 0.63. The QP estimates for household characteristics seem very reasonable. The estimated total number of households is 1.8% larger than the census count while the A.C.E. estimated number of persons is 2.0% larger than the census count. The quadratic programming total number of persons differs slightly from the A.C.E. estimate because of rounding of the weights. The difference is about 7% of the standard error.

Table 4
The Census Count, A.C.E. Estimates and QP Estimates with A.C.E. Controls – WA 1788

	Census Count	A.C.E. Count	QPG Count (%)	s.e. (QPG) Count (%)
Housing unit characteristics				
With Own Children	4,349	—	101.89	2.09
Not With Own Children	3,685	—	101.66	3.07
With 1 to 4 Persons	6,785	—	102.03	2.03
With 5+ Persons	1,249	—	100.40	5.92
Rented Unit	2,559	—	104.57	2.62
Owned Unit	5,475	—	100.47	1.50
Total	8,034	—	101.78	1.22
Person characteristics				
Age 0–4 years	2,493	103.17	102.81	1.00
Age 5–17 years	6,339	103.09	103.08	0.96
Age 18–44 years	12,711	101.67	101.67	0.63
Age 45–64 years	3,028	100.26	100.33	0.59
Age 65+ years	574	99.48	98.95	0.70
Males	12,473	102.18	102.01	0.68
Females	12,672	101.74	101.82	0.62
Hispanic	2,385	104.95	104.91	1.09
Not Hispanic	22,760	101.64	101.60	0.60
Black	1,285	104.59	104.82	1.01
White	22,372	101.69	101.69	0.61
Asian	257	100.00	101.95	1.95
Remainder	1,231	104.47	102.92	1.14
In Rented Unit	7,978	104.25	104.21	0.89
In Owned Unit	17,167	100.89	100.84	0.68
Total	25,145	101.96	101.91	0.57

5. CONCLUSIONS

The QP method is shown to work well on actual USCB long form data. The QP single household weight method possesses several advantages over the USCB separate weights method. With one set of weights, there will be no confusion as to which weights to use for estimating a given characteristic. Also, estimates of relationships such as ratios of person characteristics to household characteristics are

expected to be less variable when a single set of weights is used for both characteristics.

Given that a single set of weights is easier to compute and easier for analysts to use, one would only construct two sets of weights if the weights designed for one type of characteristic give estimates with smaller variance for that type of characteristic. This did not seem to be the case in our example. The single set of QP weights gave favorable

results for both household and person characteristics when compared with the USCB weights for the specific category.

The QP estimation module is computationally feasible and can replace the raking estimation module in the USCB operational setting. The QP method can produce long form sample weights for households in an adjustment situation in which only person controls are available.

ACKNOWLEDGEMENTS

This article reports the results of research and analysis undertaken by the authors. It has undergone a more limited review than official U.S. Census Bureau publications. Research results and conclusions expressed are those of the authors and have not been endorsed by the U.S. Census Bureau. The report is released to inform interested parties of research and to encourage discussion.

This research was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of the Census.

We gratefully acknowledge the comments of the Associate Editor and two referees which led to a much improved paper.

APPENDIX

Procedure used to define cells and initial weights $W_i^{(2)}$

We used the USCB's procedure to determine the order in which cells are combined (collapsed). The cell collapsing rules specify that each cell contain at least 5 sample households. The procedure below is our extension of the USCB rules for defining $W_i^{(2)}$.

Let two cells under consideration be identified as Cell 1 and Cell 2.

i) Cell 1 is not to be collapsed and $n_1^{-1}N_1 \leq B$, where N_1 is the Census count of households in Cell 1 and n_1 is the long form sample count in Cell 1. The constant B is provided by the sponsor and in our work, 27 is used. For household i in Cell 1, let

$$W_i^{(2)} = \max \{1.2, \ddot{W}_i\}, \tag{A.1}$$

where $\ddot{W}_i = \min \{Q_1 W_i^{(1)}, B\}$,

$$Q_1 = \left[\sum_{i \in A_1} W_i^{(1)} \right]^{-1} N_1,$$

and A_1 is the set of indices in Cell 1. The number 1.2 is an arbitrary lower bound chosen greater than one and less than the minimum of $W_i^{(1)}$ which is two. Note that the $W_i^{(2)}$ provides reasonable estimated totals for Cell 1. If $n_1^{-1}N_1 > B$, collapse cell 1 with cell 2 as in ii) below.

ii) Cells 1 and 2 are designated for collapse, $(n_1 + n_2)^{-1}(N_1 + N_2) \leq B$, $n_1 + n_2 \geq 5$, and $n_1^{-1}N_1 > n_2^{-1}N_2$. Then for i in Cell 1, $W_i^{(2)}$ is defined by (A.1). For i in Cell 2,

$$W_i^{(2)} = \max \{1.2, \ddot{W}_i\},$$

where

$$\ddot{W}_i = \min \{Q_2 W_i^{(1)}, B\},$$

$$Q_2 = \left[\sum_{i \in A_2} W_i^{(1)} \right]^{-1} (N_1 + N_2 - \hat{N}_1),$$

and

$$\hat{N}_1 = \sum_{i \in A_1} W_i^{(2)}.$$

The $W_i^{(2)}$ in $A_1 \cup A_2$, the union of cells 1 and 2, maintains the total households in $A_1 \cup A_2$ and also provide an estimated total for Cell 1 that is reasonably close to the true total.

iii) Cells 1 and 2 are designated for collapse, $n_1 + n_2 \geq 5$, and $(n_1 + n_2)^{-1}(N_1 + N_2) > B$. Then it is necessary to initiate further collapsing. The combined cell becomes the Cell 1 of case (ii). Continue cell collapsing until $(n_1 + n_2 + ..)^{-1}(N_1 + N_2 + ..) \leq B$. Case (iii) was not observed in the study data set.

One could repeat the weight construction procedure in an iterative manner by using the $W_i^{(2)}$ as $W_i^{(1)}$ in a second cycle. We tried a second cycle on the data described in the text. There was no discernable improvement in the estimates from using a second cycle.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*. 13, 183-198.

BANKIER, M.D., RATHWELL, S. and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. Working Paper-Methodology Branch, Census operations section, Social Survey Methods Division, Statistics Canada. SSMD92-007E.

- BANKIER, M., HOULE, A.M. and LUC, M. (1997). Calibration estimation in the 1991 and 1996 canadian censuses. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 66-75.
- BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*. 33, 290-299.
- DEVILLE, J., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*. 87, 376-382.
- FULLER, W.A. (1998). Replication variance estimation for two phase samples, *Statistica Sinica*. 8, 1153-1164.
- FULLER, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*. 28, 5-23.
- FULLER, W.A., and ISAKI, C.T. (2001). Estimation using estimated coverage in a census. Presented at the CAESAR conference, June, Rome, Italy.
- FULLER, W.A., LAUGHIN, M.M. and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*. 20, 75-85.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association*. 88, 1047-1060.
- HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. Thesis, Iowa State University, Ames, Iowa.
- ISAKI, C.T., TSAY, J.H. and FULLER, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology*. 26, 31-42.
- ISAKI, C.T., IKEDA, M.M., TSAY, J.H. and FULLER, W.A. (2000). An estimation file that incorporates auxiliary information. *Journal of Official Statistics*. 16, 155-172.
- JAYASURIYA, B.R., and VALLIANT, R. (1996). An application of restricted regression estimation in a household survey. *Survey Methodology*. 22, 127-137.
- LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*. 13, 199-207.
- RAO, J.N.K., and SINGH, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 57-65.
- SCHINDLER, E., GRIFFIN, R. and SWAN, C. (1992). Weighting the 1990 census sample. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 664-669.

Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism

D. NASCIMENTO DA SILVA and JEAN D. OPSOMER¹

ABSTRACT

The weighting cell estimator corrects for unit nonresponse by dividing the sample into homogeneous groups (cells) and applying a ratio correction to the respondents within each cell. Previous studies of the statistical properties of weighting cell estimators have assumed that these cells correspond to known population cells with homogeneous characteristics. In this article, we study the properties of the weighting cell estimator under a response probability model that does not require correct specification of homogeneous population cells. Instead, we assume that the response probabilities are a smooth but otherwise unspecified function of a known auxiliary variable. Under this more general model, we study the robustness of the weighting cell estimator against model misspecification. We show that, even when the population cells are unknown, the estimator is consistent with respect to the sampling design and the response model. We describe the effect of the number of weighting cells on the asymptotic properties of the estimator. Simulation experiments explore the finite sample properties of the estimator. We conclude with some guidance on how to select the size and number of cells for practical implementation of weighting cell estimation when those cells cannot be specified *a priori*.

KEY WORDS: Finite population asymptotics; Quasi-randomization inference; Weighting cell selection.

1. INTRODUCTION

Item and unit nonresponse occur in almost all large-scale surveys, and proper estimation techniques need to account for it. While item nonresponse is often dealt with through imputation, unit nonresponse is most often accounted for through weighting adjustments. Cell weighting adjustments for nonresponse have been applied since at least the 1950s in survey estimation, *e.g.* U.S. Bureau of the Census (1963, page 53), and continue to be widely used in practice today, because they have intuitive appeal and are relatively easy to implement in practice. Reviews of common weighting procedures are given in Kalton (1983) and Kalton and Kasprzyk (1986). A number of authors have studied the properties of the weighting cell estimator under a variety of theoretical frameworks. Oh and Scheuren (1983) derive the mean and variance of the weighting cell estimator under simple random sampling, conditional on the sample size and the number of respondents in each cell. See also Kalton and Maligalig (1991). Särndal, Swensson and Wretman (1992, page 578) use the term “response homogeneity group” for cells in which the nonresponse is assumed to be constant, and derive the properties of the resulting weighting cell estimator for general designs. The recently introduced *fully efficient fractional imputation* (FEFI) of Kim and Fuller (1999) can also be expressed as a weighting cell estimator, and these authors derive its model properties under the assumption that the variables are independent and identically distributed (iid) within each cell.

While the specific assumptions vary, a common thread among all these results is that the weighting cells are correctly specified, in the sense that units within each cell are indeed fully “exchangeable” (the precise definition of this term depends on the framework selected: equal response probabilities for randomization-based inference, or iid observations for model-based inference). In the terminology of Little and Rubin (2002, Chapter 1), this is the case of observations *missing at random* (MAR), where auxiliary information (*i.e.*, cell membership in this case) can be used to correct the inference for the nonresponse.

In this article, we depart from this framework. We will assume that the response mechanism depends on a known continuous auxiliary variable, but the exact functional form of this relationship is left almost completely unspecified (details on this *nonparametric response mechanism* are provided in the next section). Knowledge of such a variable could be used to construct more sophisticated nonresponse adjustments such as *propensity weighting* (Cassel, Särndal and Wretman (1983), Little (1986), and Da Silva and Opsomer (2003)) or post-stratification, but we will instead limit our use of this auxiliary variable to the division of the population into weighting cells. Our primary goal with this approach is to study the robustness of the popular weighting cell estimator to model misspecification, and in particular, the effect of the number of cells. Hence, in contrast to the approach of the authors discussed above, the weighting cells are used as a practical way to construct an survey estimator, but they will not be assumed as part of the

¹ D. Nascimento Da Silva, Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN 59072-970, Brazil. E-mail: damiao@cet.ufrr.br; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames IA 50011, U.S.A. E-mail: jopsomer@iastate.edu.

statistical framework. This is similar to the “adjustment by subclassification” idea proposed by Cochran (1968) for removing the bias due to a continuous covariate in observational studies.

We will study the properties of the estimator under *quasi-randomization*, a term used by Oh and Scheuren (1983) to denote joint inference under the sampling design and the response mechanism. The asymptotic properties of the estimator will be established by embedding the finite population and the corresponding sampling design and response mechanism in a sequence of such populations and random mechanisms, as will be explained in later sections. This asymptotic framework is very similar to that advocated by Hansen, Madow and Tepping (1983) and used in Isaki and Fuller (1982), among others.

The remainder of this paper is as follows. In section 2, we introduce the notation and framework for the sampling design and the nonresponse model, and discuss the weighting cell estimator. In the following section, we derive the asymptotic design properties of the estimator. In section 4, we report on a simulation study to examine the practical behavior of the estimator, compare its practical behavior with that predicted by the asymptotic theory, and provide some guidance on the choice of the weighting cells.

2. THE WEIGHTING CELL ESTIMATOR

Before describing the weighting cell estimator, we introduce our survey design framework and the response generating mechanism. We consider a population $U = \{1, 2, \dots, N\}$, where N is finite and known. For every element i in U , let $\mathbf{Y}_i = (Y_{1,i}, Y_{2,i}, \dots, Y_{p,i})$ be the associated vector of values of p characteristics of interest, Y_1, Y_2, \dots, Y_p . Likewise, let $\mathbf{X}_i = (X_{1,i}, X_{2,i}, \dots, X_{q,i})$ be the vector of values of q auxiliary variables, X_1, X_2, \dots, X_q , corresponding to the i th unit, $i \in U$. We assume that \mathbf{X}_i is known $\forall i \in U$. If $p = 1$, we denote \mathbf{Y}_i by Y_i and, for $q = 1$, \mathbf{X}_i is used to denote X_i . Let s represent a sample drawn from U according to some sampling design $p(\cdot)$. This sampling design $p(\cdot)$ is chosen by the survey sampler and may be based on information available in the \mathbf{X}_i , $i \in U$.

The goal of the sample survey is to estimate unknown population quantities such as the population mean or total, or a function of these quantities. To simplify the presentation, we will focus on the estimation of the population total of the \mathbf{Y}_i ,

$$t_y = \sum_U \mathbf{Y}_i.$$

When there is no nonresponse, this quantity will be estimated by a sample-based estimator of the form

$$\hat{t}_y = \sum_s w_i \mathbf{Y}_i = \sum_U w_i \mathbf{Y}_i I_i \quad (1)$$

where the w_i , $i \in s$, are the sampling weights and I_i is an indicator for whether the i th unit is in the sample or not. In this article, we will assume that the sampling weights are the inverse of the inclusion probabilities, or $w_i = \pi_i^{-1}$, with $\pi_i = \Pr(i \in s)$, so that the estimator (1) is the classical Horvitz-Thompson estimator (Horvitz and Thompson 1952). Also, let $\mathbf{I} = (I_1, I_2, \dots, I_N)^T$ represent the vector of inclusion indicators for the population.

In the context of nonresponse, it is convenient to assume that each unit in the population is either a *respondent* or a *nonrespondent* for the variable of interest \mathbf{Y} . Consider the vector $\mathbf{R} = (R_1, R_2, \dots, R_N)^T$, where R_i indicates if the i th unit is a respondent or not. The distribution of \mathbf{R} is called the *response mechanism*. In analogy to the definition of the sample s , we use $r \subseteq U$ to denote the (realized) set of respondents in the population, i.e., those elements for which $R_i = 1$. Since the distribution of r and \mathbf{R} is typically unknown and can in principle depend on the realized value of \mathbf{I} as well as on the \mathbf{Y} , we need to assume a model for the response mechanism. When this assumed model is used to develop an estimator for a population quantity, the properties of this estimator become dependent on the response model. Hence, a misspecified model for \mathbf{R} has the potential to cause significant and difficult to measure bias in both the estimator and its associated measures of precision. To avoid this problem, we will keep the response mechanism quite general in this article. Specifically, we will assume that the R_i are independent Bernoulli variables with

$$\Pr\{R_i = 1 | \mathbf{I}, \mathbf{Y}\} = \varphi_i, \quad 0 < \varphi_i \leq 1, \quad \forall i \in U,$$

and that the φ_i can be written as $\varphi_i = \varphi(\mathbf{X}_i)$, with $\varphi(\cdot)$ a continuous and differentiable but otherwise unspecified function of the \mathbf{X}_i . Note that this includes the uniform response mechanism, where $\varphi_i \equiv \varphi$ for all $i \in U$, as a special case.

When some of the selected elements do not respond, the estimator (1) can no longer be computed, and an estimator that includes a nonresponse adjustment is required. In this article, we are using the weighting cell estimator for this purpose. For simplicity, we will describe the situation in which both the \mathbf{Y}_i and \mathbf{X}_i are univariate variables, but the approach can be generalized to the multi-dimensional case. Let $s_r = s \cap r$ represent the subset of the selected elements that actually respond to the survey.

Let U_g , $g = 1, \dots, G$, represent G groups obtained by dividing the population into groups based on the values of the known auxiliary variable X . Specific implementations might generate groups of equal size, or divide the range of

X into equal-length intervals. We shall leave the implementation unspecified for now, and state some general assumptions about G and the size of the groups in the next section. Note that we are considering the groups as fixed with respect to the sampling design and the response mechanism, which excludes the situation in which groups are formed based on the *observed* sample values $\{X_i : i \in s\}$. This was done primarily to simplify the theoretical derivations, and is similar to the approach of Särndal *et al.* (1992) and Kim and Fuller (1999), among others.

Let $s_g = s \cap U_g$ be the portion of the sample that falls in group g , and define similarly $s_{r,g} = s_r \cap U_g$. The weighting cell estimator is defined as

$$\hat{t}_{WC} = \sum_{g=1}^G \left(\frac{\sum_{s_g} w_i}{\sum_{s_{r,g}} w_i} \right) \sum_{i \in s_{r,g}} w_i Y_i. \quad (2)$$

From this expression, it is easy to see that in each group, the estimator of the group total is ratio-adjusted by the inverse of the weighted proportion of respondents in the cell. This estimator is also the FEFI estimator of Kim and Fuller (1999). The properties of this estimator will be studied in next section.

3. PROPERTIES UNDER QUASI-RANDOMIZATION

3.1 Asymptotic Framework and Assumptions

The quasi-randomization properties of the weighting cell estimator will be studied in the usual finite population asymptotic context, in which the population U is treated as an element in an increasing sequence U_1, U_2, \dots, U_v with $v \rightarrow \infty$, with a corresponding sequence of sampling designs $p_v(\cdot)$ (see Isaki and Fuller (1982) for an early example of this framework). Let N_v be the size of the v^{th} population with $N_v > N_{v-1}$, let $Y_v = (Y_1, Y_2, \dots, Y_{N_v})^T$ denote the set of values of the characteristic of interest, Y , associated with U_v , and similarly, $X_v = (X_1, X_2, \dots, X_{N_v})^T$. We assume that X_v is known. For each v , a sample of size n_v ($n_v \geq n_{v-1}$) is selected from U_v , according to a sampling design $p_v(\cdot)$. As before, let $I_v = (I_1, I_2, \dots, I_{N_v})^T$ be the corresponding sample inclusion vector. We will denote the K^{th} order central moment of the sample membership indicators I_{i_1}, \dots, I_{i_K} by

$$\Delta_{i_1, \dots, i_K} = E \left(\prod_{k=1}^K (I_{i_k} - \pi_{i_k}) \right). \quad (3)$$

It is assumed that U_v can be divided into G_v ($G_v \geq G_{v-1}$) mutually exclusive and exhaustive groups, U_g , $g = 1, \dots, G_v$. These groups are constructed by sorting the

population according to their X values and dividing the population into G_v groups. We will assume that there are at least G_v distinct values among the elements of X_v . Let N_g represent the number of elements in U_g .

As mentioned in the previous section, we are treating the groups as fixed with respect to the population. The problem created by this approach is that in general, there is a non-zero chance of obtaining a group without any respondents. We solve this problem by adding a small constant in the denominators in each of the groups, or

$$\hat{t}_{WC}^* = \sum_{g=1}^{G_v} \left(\frac{\sum_{s_g} w_i}{\max \left(\sum_{s_{r,g}} w_i, N_g G_v n_v^{-1} \right)} \right) \sum_{i \in s_{r,g}} w_i Y_i. \quad (4)$$

Hence, the difference between \hat{t}_{WC}^* and \hat{t}_{WC} in (2) is asymptotically negligible. This is similar to what is often done in practice to avoid overly large weights in ratio estimation.

Fuller and Kim (2003) give the limiting distribution of the FEFI estimator under the assumption that the response probabilities are constant within these cells. We will study the case where the response probabilities are a smooth function of an auxiliary variable and the number of cells are allowed to vary. Let $\mathbf{R}_v = (R_1, R_2, \dots, R_{N_v})^T$ be the response indicator vector for the v^{th} population. We assume that the distribution of \mathbf{R}_v satisfies the *nonparametric response mechanism* assumptions, specified as follows:

- (R1) R_1, R_2, \dots, R_{N_v} are independent random variables,
- (R2) $\Pr \{ R_i = 1 \mid \mathbf{I}_v, \mathbf{Y}_v \} = \phi_i, \forall i \in U_v$,
- (R3) $\phi_i = \phi(X_i) \forall i \in U_v$, where $\phi(\cdot)$ is differentiable with bounded first derivative, and the $X_i \in [x_m, x_M]$, with x_m, x_M fixed constants and $x_m < x_M$.

The remaining assumptions are technical conditions that will be used extensively in the proofs. We assume that there are positive constants $\lambda_1, \lambda_2, \dots, \lambda_9$ such that:

- (A1) $\lambda_1 < N_v n_v^{-1} \pi_i < \lambda_2 < \infty, \forall i \in U_v$, and $n_v N_v^{-1} \rightarrow \pi \in (0, 1)$, as $v \rightarrow \infty$;
- (A2) For distinct $i_1, \dots, i_K \in U_v, K = 2, 3, \dots, 8$,

$$|\Delta_{i_1, \dots, i_K}| \leq \begin{cases} \left(\prod_{k=1}^K (N - k + 1) \right)^{-1} n_v^{K/2} \lambda_3, & \text{if } K \text{ is even} \\ \left(\prod_{k=1}^K (N - k + 1) \right)^{-1} n_v^{(K-1)/2} \lambda_4, & \text{if } K \text{ is odd} \end{cases}$$
- (A3) $\lim_{v \rightarrow \infty} \frac{1}{N_g} \sum_{i \in U_g} \phi_i = \phi_g^*, \forall g = 1, 2, \dots, G_v$ and $v \geq 1$;
- (A4) $\max_{i \in U_v} |Y_i| \leq \lambda_5$;
- (A5) $\lambda_6 < \min_{i \in U_v} \phi_i \leq 1$;

$$(A6) \quad \lambda_{\gamma} G_v^{-1} \leq N_g N_v^{-1} \leq \lambda_g G_v^{-1}, \forall g = 1, 2, \dots, G_v;$$

$$(A7) \quad 1 \leq G_v \leq n_v^{\gamma} \lambda_g, \text{ with } 0 \leq \gamma \leq 1/2.$$

Assumptions (A1) – (A2) imply that, asymptotically, the sampling design is “well behaved,” in the sense that the moments of the sample membership indicators are of the same order of magnitude as those in simple random sampling without replacement. This is a common assumption in finite population asymptotic theory. (A1) also requires that the sampling fraction converges to a constant in the interval (0, 1). The boundedness assumption (A4) on the observations will significantly simplify the proofs for some of the theorems in the article, and could be relaxed to the existence of bounded moments if desired. Similarly, some technical regularity conditions are required to avoid degenerate response mechanisms: (A3) provides that the limit for the average response probability in a cell exists, and (A5) excludes the situation in which some units might have $\phi_i = 0$. Finally, assumptions (A6) and (A7) on the weighting cells require that all the cells grow at a similar rate, and that the total number of cells does not increase “too fast” relative to the sample size.

3.2 Main Results

The approach we will use in the study of the properties of the weighting cell estimator follows that commonly used in the study of finite population estimators. First, we show the asymptotic equivalence between the non-linear weighting cell estimator and a “linearized” approximation. Next, we derive the mean squared error properties of the linearized estimator and consider those as the asymptotic properties of the weighting cell estimator or, more precisely, the properties of the asymptotic distribution of the weighting cell estimator. See, for instance, Särndal *et al.* (1992, Chapter 5) for a description of this approach.

The following theorem formally states our first results. The proof is in the appendix.

Theorem 3.1. *Consider the sequence of populations $\{U_v: v \geq 1\}$. Assume that for each v , a probabilistic sample of fixed size $n_v (n_v \geq n_{v-1})$ is selected from U_v according to sampling design $p_v(\cdot)$, and that the response mechanism satisfies the conditions (R1) – (R2). Finally, assume that (A1) – (A7) hold. Then, the estimator \hat{t}_{WC}^* is asymptotically equivalent to a linearized random variable \tilde{t}_{WC} , in the sense that*

$$\frac{1}{N_v} (\hat{t}_{WC}^* - \tilde{t}_{WC}) = O_p(G_v n_v^{-1}). \quad (5)$$

The bias and variance of \tilde{t}_{WC}/N_v are given by

$$E\left(\frac{\tilde{t}_{WC}}{N_v}\right) - \bar{Y}_v = \frac{1}{N_v} \sum_{g=1}^{G_v} \sum_{U_g} \left(\frac{\phi_i - \bar{\phi}_g}{\bar{\phi}_g}\right) (Y_i - \tilde{Y}_g) \quad (6)$$

and

$$\begin{aligned} \text{Var}\left(\frac{\tilde{t}_{WC}}{N_v}\right) &= \frac{1}{N_v^2} \sum_{g=1}^{G_v} \sum_{g'=1}^{G_v} \left[\sum_{U_g} \sum_{U_{g'}} \Delta_{gg'} \tilde{Y}_{ig} \tilde{Y}_{jg'} \right] \\ &\quad + \frac{1}{N_b^2} \sum_{g=1}^{G_v} \sum_{U_g} \pi_i^{-2} \frac{\phi_i(1-\phi_i)}{\bar{\phi}_g^2} (Y_i - \tilde{Y}_g)^2, \quad (7) \end{aligned}$$

where

$$\bar{\phi}_g = \frac{1}{N_g} \sum_{U_g} \phi_i, \quad \bar{Y}_g = \frac{1}{N_g} \sum_{U_g} Y_i, \quad \tilde{Y}_g = \frac{\sum_{U_g} \phi_i Y_i}{\sum_{U_g} \phi_i}$$

and

$$\tilde{Y}_{ig} = \frac{\phi_i(Y_i - \tilde{Y}_g) + \bar{\phi}_g \tilde{Y}_g}{\pi_i \bar{\phi}_g}, \quad \forall i \in U_g \text{ and } \forall g = 1, 2, \dots, G_v.$$

Remark 1. The asymptotic equivalence between \hat{t}_{WC}^* and \tilde{t}_{WC} depends on the number of groups G_v , with a faster convergence rate achieved when G_v grows more slowly. The intuition behind this result is that the goodness of the linear approximation depends on how well the true cell ratio response adjustments ϕ_g^* are estimated by the sample-based estimators $\sum_{s_g} w_i / \sum_{s_g} w_i$. Since the cell ratios will be better estimators as the sample size grows larger, this would argue that G_v should be chosen to be small, which corresponds to the current practice in applications of weighting cell estimation. However, as will be shown below, the MSE properties of \hat{t}_{WC} under the nonparametric response mechanism improve as G_v gets larger. A more detailed discussion of the selection of the number of groups will be provided after Theorem 3.2 below and in section 4.

Remark 2. The results in Theorem 3.1 depend on the population groups $U_g, g = 1, \dots, G_v$ and on the $\phi_i, i \in U_v$, but do not rely on the fact that the response probabilities are a smooth function of the auxiliary variable X . Hence, the explicit expressions for the asymptotic bias and variance can be used to derive results for other response mechanisms that follow (R1) – (R2). In particular, results for the response homogeneity group model (see Särndal *et al.* 1992, page 577) follow directly from Theorem 3.1. This is also the model studied by Fuller and Kim (2003). Under that model, one assumes that $\phi_i \equiv \phi_g$ for all

$i \in U_g, g = 1, \dots, G$, and it can easily be shown that the bias of \tilde{t}_{wc} is 0 and its variance is

$$\text{Var}\left(\frac{\tilde{t}_{wc}}{N_v}\right) = \text{Var}\left(\frac{\hat{t}_y}{N_v}\right) + \frac{1}{N_v} \sum_{g=1}^G \frac{1 - \phi_g}{\phi_g} \sum_{U_g} \pi_i^{-2} (Y_i - \bar{Y}_g)^2.$$

The first term in the variance is the variance of the estimator without nonresponse, and the second term represents the variance inflation caused by the nonresponse under a homogeneous within-cell response mechanism.

The following corollary follows directly from Theorem 3.1 and Fuller (1996, Theorem 5.2.1). A proof is given in the appendix.

Corollary 3.1. *Under the conditions of Theorem 3.1 with $\gamma < 1/2$ in (A7), for any sampling design $p_v(\cdot)$ such that*

$$n_v^{1/2} \left(\frac{\tilde{t}_{wc}}{N_v} - \bar{Y}_v - B_v \right) \xrightarrow{L} N(0, V),$$

where B_v corresponding to the bias of \tilde{t}_{wc}/N_v given in Theorem 3.1 and

$$V \equiv \lim_{v \rightarrow \infty} n_v \text{Var}(\tilde{t}_{wc}/N_v) \in (0, \infty),$$

then

$$\left[\text{Var}\left(\frac{\tilde{t}_{wc}}{N_v}\right) \right]^{-1/2} \left(\frac{\hat{t}_{wc}^*}{N_v} - \bar{Y}_v - B_v \right) \xrightarrow{L} N(0, 1).$$

Corollary 3.1 states that, whenever the linearized estimator \tilde{t}_{wc} achieves asymptotic normality, then so does \hat{t}_{wc}^* . Since \tilde{t}_{wc} can be written as a classical expansion estimator of the form (1), this result is quite general.

Under the nonparametric response mechanism described in (R1) – (R3), it is possible to describe the effect of the number of groups G_v on the asymptotic bias and variance of \hat{t}_{wc}^* . The next theorem gives the asymptotic rates for the bias and variance, and is proven in the appendix.

Theorem 3.2. *Assume that (R3) and the conditions of Theorem 3.1. Then,*

$$E\left(\frac{\tilde{t}_{wc}}{N_v}\right) - \bar{Y}_v = O\left(\frac{1}{G_v}\right)$$

and

$$\text{Var}\left(\frac{\tilde{t}_{wc}}{N_v}\right) = O\left(\frac{1}{n_v}\right) + O\left(\frac{1}{n_v G_v}\right).$$

Remark 3. Theorem 3.2 shows that both the asymptotic bias and variance of the weighting cell estimator \hat{t}_{wc}^* become smaller as the number of groups G_v increases. An intuitive explanation of that fact is that the approximation of the function $\phi_i = \phi(X_i)$ by the step function $\phi_i = \phi_g^*$ improves as the number of cells increases. The asymptotic variance has a term that is independent of G_v . This “residual variance” is due to the inherent variability of the sampling design and the response mechanism, and cannot be reduced by changing G_v .

Remark 4. As noted in Remark 1, constructing a good linear approximation \tilde{t}_{wc} requires G_v to be small, while Theorem 3.2 states that the MSE of \tilde{t}_{wc} is minimized by taking G_v as large as possible. Taken together, this can be interpreted to mean that, once the sample size in every cell is sufficiently large to obtain a “valid” ratio estimator for the average cell response probability ϕ_g^* , it is preferable to increase the number of cells than to increase the sample size per cell. The simulation experiments discussed in section 4 will further explore this recommendation.

The following corollary follows directly from Corollary 3.1, Theorem 3.2, and Chebyshev’s inequality, and establishes the consistency of the weighting cell estimator under the nonparametric response mechanism.

Corollary 3.2. *Under the conditions of Theorem 3.2, \hat{t}_{wc}^* is a consistent estimator for t_y , in the sense that for any $\epsilon > 0$,*

$$\Pr\left(\left|\frac{\hat{t}_{wc}^* - t_y}{N_v}\right| > \epsilon\right) \rightarrow 0, \quad v \rightarrow \infty.$$

Remark 5. As Corollary 3.2 shows, as long as a variable X can be found that is sufficiently related to the nonresponse, in the sense of assumptions (R1) – (R3), construction of weighting cells does not require knowledge of homogeneous response probability cells in order to construct a consistent estimator. However, as discussed in Remarks 1 and 4, the choice of the number of cells still has an effect on the properties of the estimator.

Remark 6. Assumption (R3) can easily be relaxed to allow for a small number of points of discontinuity in both $\phi(\cdot)$ and its first derivative. A “small” number can mean that the number is either fixed as $v \rightarrow \infty$ or increases at a rate slower than G_v . This would make it possible to account for situations such as stratified designs or the presence of domains within U_v . The present theory can be extended

directly to these situations, if the values for the variable X fall in non-overlapping segments for the different strata or domains.

4. SIMULATION EXPERIMENTS

4.1 Description of the Experiment

In order to investigate the practical implications of the results of section 3, we carried out a Monte Carlo experiment on a fixed population of $N = 3,000$ units. We consider the case of one covariate, X , whose population values are generated as:

$$X_1, X_2, \dots, X_N \sim \text{i.i.d. } U(0, 1),$$

and two different variables of interest, Y_1 and Y_2 . We are interested in evaluating the effects of (1) the (model) relationship between Y and X , (2) the response mechanism $\varphi(X)$, (3) the sample size n and (4) the number of cells G , on the bias and on the mean square error of the \hat{t}_{WC} estimator. Since our theoretical results rely on the approximation of \hat{t}_{WC} (or \hat{t}_{WC}^*) by a linearized estimator \tilde{t}_{WC} , we will also compare the behavior of \hat{t}_{WC}/N_v and \tilde{t}_{WC}/N_v as estimators of the population mean, $\bar{Y}_v = N_v^{-1} \sum_U Y_i$. Finally, we compare \hat{t}_{WC}/N_v to the “naive” estimator of the mean, which is defined for the variable Y as:

$$\bar{y}_r = \frac{\sum_{i \in s_r} w_i Y_i}{\sum_{i \in s_r} w_i},$$

corresponding to a ratio adjustment of the respondent sample to the original sample. This estimator is appropriate under the assumption of uniform response mechanism or, to use the terminology of Little and Rubin (2002, chapter 1), when observations are *missing completely at random* (MCAR). Note that \bar{y}_r is equivalent to the weighting cell estimator with a single cell.

The levels of the four factors used in the experiment are given in Table 1. The “levels” of the variable Y correspond to two populations of independent values. The variable Y_1 was generated as $N(40, 58)$, truncated to -3 to +3 standard deviations, corresponding to the “white noise” case. The variable Y_2 is related to X and was generated through the linear model $Y_2 = 27.12 + 26.06X + \varepsilon$, where $\varepsilon \sim N(0, 9)$. The population mean and variance for the two variables were, respectively, (39.9, 55.3) for Y_1 , and (40.0, 63.9) for Y_2 .

The four levels of the response mechanisms contain two different scenarios regarding the response probabilities: constant ($C1$, $C2$), and linearly related to X ($L1$, $L2$). The response probabilities are:

- $\varphi_{C1}(X) = 0.5$
- $\varphi_{C2}(X) = 0.8$
- $\varphi_{L1}(X) = 0.20 + 0.60X$
- $\varphi_{L2}(X) = 0.65 + 0.30X$

The levels of the linear response mechanisms were chosen so that the average probabilities (over X) were approximately equal to 0.5 and 0.8, respectively.

Table 1
Overview of Factors in the Simulation Experiment

Factor	Levels
Y variable	Y_1, Y_2
Response mechanism $\varphi(\cdot)$	$C1, C2, L1, L2$
Sample size n	200, 500
Number of cells G	2, 3, 5, 8

For a given G , the groups were created by dividing the range of X into G equal segments and assigning the element i to the group g if the value X_i was in the g^{th} segment, $i = 1, 2, \dots, N$ and $g = 1, 2, \dots, G$. The simulations were carried out through a completely randomized factorial experiment $2 \times 4 \times 2 \times 4$. For each combination of the levels of the factors in Table 1, $B = 5,000$ independent realizations of the vector indicator of responses, $\mathbf{R} = (R_1, R_2, \dots, R_N)^T$, were generated according to the corresponding response mechanism. For each one of such realizations, a simple random sample (without replacement and of size n), s , was selected from the overall population. Within each selected sample, the respondents were the values of $i \in s$ such that $R_i = 1$.

This procedure could in principle lead to a group not containing any sampled and responding element, in which case the weighting cell estimator (ignoring the adjustment in (4)) cannot be computed. If that happened, the realization was discarded and a new sample drawn from the population. Out of the 5,000 repetitions for each combination of factors, this happened 13 times in the factor combination ($Y_1, \varphi_{L1}, 200, 8$) and 15 times with ($Y_2, \varphi_{L1}, 200, 8$). It did not occur with any of the other factor combinations. Hence, the number of samples discarded was very small and this has a negligible effect on the simulation results.

With $n = 200$ and $G = 8$, we expect approximately 25 sampled elements in each cell, to be further reduced by the nonresponse. Since the estimator relies on ratio estimation in each cell, we judged this to be a reasonable lower bound on the number of observations per cell to consider in the simulations. In practice, a number of procedures could be used when groups have too few elements, such as picking

a smaller value for G or collapsing neighboring groups. We also implemented an estimator that collapses the empty cell with a neighboring cell as well as a version with a lower bound on the value of the denominator in the weighting adjustment (i.e., \hat{t}_{WC}^*), and the results are virtually indistinguishable from those reported below, so they will not be further discussed here.

4.2 Results

Tables 2 and 3 show the simulated bias of the weighting cell estimator for the variables Y_1 and Y_2 as a fraction of the standard deviation. As a comparison, the last column of Tables 2 and 3 displays the bias of the naive estimator, \bar{y}_r . The bias as a fraction of the standard deviation, referred to here as the *relative bias*,

$$RB(\hat{t}_{WC}, \hat{t}_y) = \frac{E(\hat{t}_{WC} - \bar{Y})}{(\text{Var}(\hat{t}_{WC}))^{1/2}}$$

was also used in Cochran (1977, page 14), where it is shown that as the relative bias increases, inferential results rapidly become unreliable. In a simple simulation example, Cochran (1977) shows that a relative bias of ± 0.50 or more leads to highly inaccurate 95% confidence intervals.

For Y_1 (Table 2), the relative bias of the weighting cell estimator is small and is similar to the relative bias of the naive estimator, for all sample sizes, response mechanisms and cells sizes considered. For the variable Y_2 (Table 3), similar results hold when the response mechanism is uniform (C1, C2). However, when the response probabilities are a linear function of X (L1, L2), the naive estimator becomes severely biased. This relative bias decreases as the number of cells increases, and three to five cells appear sufficient to remove most of the bias. This finding agrees with that of Cochran (1968) in the context of bias reduction for observational studies.

Table 2
Relative Bias of the Weighting Cell and Naive Estimators
for the Mean Y_1

Sample size	Response mechanism	Number of Cells				Naive estimator
		2	3	5	8	
200	C1	-0.00	-0.01	0.01	0.01	-0.00
	C2	0.01	-0.00	-0.01	0.00	0.00
	L1	-0.02	0.03	-0.04	-0.01	-0.00
	L2	-0.00	-0.02	0.00	-0.02	-0.00
500	C1	-0.00	-0.01	0.04	-0.01	0.00
	C2	0.01	0.02	-0.01	-0.01	0.00
	L1	0.05	0.02	-0.01	-0.02	0.01
	L2	0.01	0.01	-0.00	-0.01	0.01

Table 3
Relative Bias of the Weighting Cell and Naive Estimators
for the Mean of Y_2

Sample size	Response mechanism	Number of Cells				Naive estimator
		2	3	5	8	
200	C1	0.01	-0.01	-0.02	0.02	-0.01
	C2	-0.03	-0.00	0.02	0.01	-0.00
	L1	1.16	0.59	0.22	0.07	3.57
	L2	0.36	0.18	0.06	0.03	1.36
500	C1	0.01	0.01	-0.02	-0.00	0.00
	C2	0.02	-0.00	-0.00	-0.01	-0.01
	L1	1.98	0.96	0.32	0.15	5.84
	L2	0.61	0.29	0.09	0.02	2.26

Hence, when the variable of interest is totally unrelated to the response mechanism, as in the cases of Y_1 under all mechanisms considered and of Y_2 under the uniform response mechanism, the bias does not depend on the number of cells. When the variable of interest and the response mechanism are related, multiple cells are required to remove the bias.

The relative mean squared error (RMSE) for the two variables of interest, defined as the MSE of the weighting cell estimator divided by the MSE of the estimator with no non-response,

$$RMSE(\hat{t}_{WC}, \hat{t}_y) = \frac{E(\hat{t}_{WC} - t_y)^2}{E(\hat{t}_y - t_y)^2},$$

are in Tables 4 and 5. In these tables, the last column again corresponds to the relative MSE of the naive estimator. Note that with the exception of the two L1 cases for variable Y_2 , the Tables 4 and 5 are really variance tables, since the bias is so small.

For Y_1 (Table 4), the variable uncorrelated with X , the number of cells has relatively little effect on the relative mean square error, with results around 2.3 for a 50% response rate, and around 1.3 for the 80% rate. However, a relatively modest increase in MSE is observed, especially for the high nonresponse cases (C1, L1). For Y_2 (Table 5), the variable correlated with X , increasing the number of cells improves the results for all response mechanisms, but the effect is much more pronounced when the response mechanism is also correlated with the variable of interest. As for the relative bias, three to five cells achieve most of the efficiency gain, while the naive estimator is extremely inefficient.

Table 4
Relative Mean Squared Error of the Weighting Cell Estimator
Compared to the Estimator Without Nonresponse for Y_1

Sample size	Response mechanism	Number of Cells				Naïve estimator
		2	3	5	8	
200	C1	2.02	2.13	2.11	2.21	2.08
	C2	1.25	1.31	1.29	1.28	1.28
	L1	2.34	2.32	2.61	2.70	2.08
	L2	1.30	1.29	1.29	1.31	1.28
500	C1	2.25	2.21	2.19	2.31	2.23
	C2	1.30	1.32	1.34	1.29	1.30
	L1	2.55	2.57	2.62	2.70	2.22
	L2	1.32	1.35	1.33	1.34	1.31

Table 5
Relative Mean Squared Error of the Weighting Cell Estimator
Relative to the Estimator Without Nonresponse for Y_2

Sample size	Response mechanism	Number of Cells				Naïve estimator
		2	3	5	8	
200	C1	1.33	1.17	1.10	1.07	2.07
	C2	1.09	1.05	1.02	1.02	1.26
	L1	3.14	1.57	1.16	1.12	26.32
	L2	1.23	1.07	1.03	1.01	3.57
500	C1	1.35	1.19	1.10	1.09	2.22
	C2	1.09	1.05	1.03	1.03	1.30
	L1	6.60	2.30	1.23	1.13	69.75
	L2	1.50	1.14	1.04	1.02	7.83

The difference between the results for both variables is surprising at first, but it can be explained using the results from section 3. Clearly, the results for Y_2 follow the asymptotic theory, in that the MSE improves as the number of cells improves (as long as sufficient observations are available in each cell). In the case of Y_1 , note first that the bias is negligible relative to the standard deviation for all values of G (see Table 2), so that the change in MSE is due almost exclusively to differences in variance. It turns out that when a variable is iid in the population and sampling is equal-probability, the asymptotic variance in Theorem 3.1 is relatively insensitive to the number of cells. In that case, the increase in MSE is influenced by the variability implied in the linear approximation in Theorem 3.1, which increases with the number of cells.

The theory described in this article applies to response functions that can have arbitrary smooth shape. In order to evaluate results for more complicated functions, we also

created a variable $Y_3 = 25 + 95X - 95X^2 + \varepsilon$, where $\varepsilon \sim N(0, 3)$, so that the Y_3 has mean 40.9 and variance 51.8, and two additional quadratic response mechanisms

$$- \phi_{Q1}(X) = 0.17 + 1.96X - 1.96X^2$$

$$- \phi_{Q2}(X) = 0.50 + 1.80X - 1.80X^2.$$

The results (not shown) broadly reflect the findings for the previous variables. When the response mechanism and the variables are correlated (the linear variable is correlated with the linear response mechanism, and the quadratic variable is correlated with the linear and quadratic response mechanisms), significant bias occurs but can be removed by increasing the number of cells. In the case of the quadratic response mechanism and the quadratic variable, eight or more cells appear to be required to remove the bias. Similarly, the relative efficiency improves for all response mechanisms for both the linear (Y_2) and quadratic variable, with the most dramatic results found for the linear variable/linear response and quadratic variable/quadratic response cases.

In the previous sections of this article, we approximated the weighting cell estimator by a “linearized” estimator \tilde{t}_{wc} , and then derived the asymptotic properties of that estimator. It is therefore of interest to compare the statistical properties of both estimators in simulated settings. For all the scenarios in Table 1, we calculated the relative efficiencies of the weighting cell estimator compared to the linearized estimator. These relative efficiencies were all close to 1.00, with the largest deviation being a value of 1.08. Hence, the statistical properties of weighting cell estimator appear to be well approximated by those of the linearized estimator.

5. CONCLUSIONS

We have shown that the weighting cell estimator, corresponding also to the FEF1 estimator proposed by Kim and Fuller (1999), is consistent with respect to the sampling design and a nonparametric response model. That model does not require the correct specification of homogeneous response probability cells, as long as a variable related to the response probability can be identified.

The statistical properties of the estimator depend on the number of cells used in the estimation, but the relationship is rather complex. Asymptotically, there appears to be a trade-off between the goodness of the approximation of the weighting cell estimator by a linearized estimator, which requires a small number of cells, and the mean squared error of that linearized estimator, which is reduced when a large number of cells are used. While useful in understanding the asymptotic behavior of the estimator, these

findings only provide limited guidance for choosing the number of cells for a particular survey. However, these findings show that reliable inference for weighting cell estimators will require cells with reasonable sample sizes, because variance estimates typically rely on the variance of the linearized estimator as an approximation of the variance of the weighting cell estimator.

The simulation experiments show that when the variable of interest and the response mechanism are uncorrelated, the number of cells has virtually no effect on the design bias of the estimator. When the variable of interest and the response mechanism are uncorrelated, even the estimator with a single weighting cell (corresponding to a simple ratio adjustment) is essentially unbiased, while models with multiple cells perform equally well. When the response mechanism and the variable of interest are related, however, the bias properties of the weighting cell estimator depend critically on the number of cells. In particular, estimators with a single cell are severely biased, but even a relatively small number of cells is sufficient to reduce both the bias and variance of the estimator. This result holds for both linear and nonlinear relationships between the response mechanism and the variable of interest.

The design efficiency of estimators depends on the relationship between the variable of interest and the variable(s) used to form weighting cells. When those two variables are uncorrelated, the number of cells has no effect on the efficiency of the estimator. Conversely, when those two variables are correlated, increasing the number of cells improves the design efficiency of the estimator. Even a small number of cells dramatically improves the performance of the estimator.

Overall, it appears that in the presence of nonresponse, forming at least a small number of weighting cells based on a variable related to the non-response provides a good “insurance policy” against design bias and design inefficiency. This article has shown that this adjustment does not require the assumption that the cells be based on a priori knowledge of constant nonresponse groups. The resulting weighting cell estimator will never perform worse than the naive estimator with a single ratio adjustment for the whole sample, and it might perform significantly better.

6. ACKNOWLEDGEMENTS

The authors thank Wayne Fuller for many helpful comments made during the development of this manuscript. We also are grateful for the comments of the associate editor and the two referees. This research was supported by a subcontract between Westat and Iowa State University under Contract No. ED-99-CO-0109 between Westat and

the U.S. Department of Education. The first author gratefully acknowledges the support of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil, during his Ph. D. studies at Iowa State University.

APPENDIX

Derivations of Theoretical Results

Lemma 1. Assume that the conditions (A1) – (A3) and (R1) – (R2) hold. For $i_1, i_2, \dots, i_k \in U_v$, define

$$\Gamma_{i_1, \dots, i_k} = E \left(\prod_{i=1}^k (I_i R_{i_i} - \pi_{i_i} \phi_{i_i}) \right),$$

where $\phi_i = \phi(X_i)$. Consider the Δ_{i_1, \dots, i_k} of (3). Let A^r denotes the r -fold Cartesian product of the set A , where r is a fixed positive integer, $A_{1, r, v} = \{(i_1, i_2, \dots, i_r) \in U_v^r : i_1 = i_2 = \dots = i_r\}$ and $A_{k, r, v} = \{(i_1, i_2, \dots, i_r) \in U_v^r : \text{exactly } k \text{ components are distinct}\}$, $k = 2, 3, \dots, r$. Then, for $r = 8$,

$$N_v^8 n_v^{-8} \max_{i_1, \dots, i_8 \in A_{k, 8, v}} (|\Gamma_{i_1, \dots, i_8}|, |\Delta_{i_1, \dots, i_8}|) = \begin{cases} O(N_v^3 n_v^{-4}), & \text{if } k=5 \\ O(N_v^3 n_v^{-5}), & \text{if } k=6 \\ O(N_v n_v^{-4}), & \text{if } k=7 \\ O(n_v^{-4}), & \text{if } k=8. \end{cases}$$

Proof of Lemma 1. See Da Silva (2003).

Lemma 2. Suppose the conditions of Theorem 3.1 hold. Consider the vectors $\hat{t}_{gv} = (\hat{t}_{1,g}, \hat{t}_{2,g}, \hat{t}_{3,g})' = \sum_{U_g} \pi_i^{-1} (1, Y_i R_i, R_i)' I_i$ and $\hat{t}_{gv}^* \equiv (\hat{t}_{1,g}^*, \hat{t}_{2,g}^*, \hat{t}_{3,g}^*)'$, with $\hat{t}_{3,g}^* = \max \{\hat{t}_{3,g}, N_g G_v / n_v\}$. Let $t_{gv} = E(\hat{t}_{gv})$. Then for all $g = 1, 2, \dots, G_v$,

$$\frac{1}{N_g^8} (E \|\hat{t}_{gv}^* - t_{gv}\|^8, E \|\hat{t}_{gv} - t_{gv}\|^8) = O((G/n_v)^4).$$

Proof of Lemma 2: See Da Silva (2003).

Proof of Theorem 3.1: Consider the proof of (5). Let $a = (a_1, a_2, a_3)' \in \mathbb{R}^3$ and $h: \mathbb{R}^3 \rightarrow \mathbb{R}$, where $h(a) = a_1 a_2 / a_3$, $a_3 \neq 0$. Define

$$\eta_{gv}(a) = h(N_g^{-1} t_{gv}) + \sum_{k=1}^3 h^{(k)}(N_g^{-1} t_{gv})(a_k - N_g^{-1} t_{gv}),$$

where $h^{(k)}(a) = \partial h(a) / \partial a_k$, and let $e_{gv} = h(a) - \eta_{gv}(a)$. Note that $\hat{t}_{wc}^* = \sum_{g=1}^{G_v} N_g h(N_g^{-1} \hat{t}_{gv}^*)$, and hence, defining the “linearized” estimator $t_{wc} = \sum_{g=1}^{G_v} N_g \eta_{gv}(N_g^{-1} \hat{t}_{gv}^*)$, we can write

$$\frac{1}{N_v} (\hat{t}_{wc}^* - t_{wc}) = \bar{e}_v + \bar{\eta}_v,$$

where

$$\bar{e}_v = \frac{1}{N_v} \sum_{g=1}^{G_v} N_g e_{gv} (N_g^{-1} \hat{t}_{gv}^*)$$

and

$$\bar{\eta}_v = \frac{1}{N_v} \sum_{g=1}^{G_v} N_g (\eta_{gv} (N_g^{-1} \hat{t}_{gv}^*) - \eta_{gv} (N_g^{-1} \hat{t}_{gv})).$$

Consider first the term $\bar{\eta}_v$. Observe that

$$\begin{aligned} |\eta_{gv} (N_g^{-1} \hat{t}_{gv}^*) - \eta_{gv} (N_g^{-1} \hat{t}_{gv})| &= \\ |h^{(3)}(N_g^{-1} \mathbf{t}_{gv})| \frac{1}{N_g} |\hat{t}_{3,g}^* - \hat{t}_{3,g}|. \end{aligned}$$

By (A4) and (A5), it is straightforward to check that $h(\cdot)$ and $h^{(k)}(\cdot)$, $k = 1, 2, 3$, are $O(1)$ when evaluated at $N_g^{-1} \mathbf{t}_{gv}$, for all $g = 1, 2, \dots, G_v$. Since by construction, we have $1/N_g |\hat{t}_{3,g}^* - \hat{t}_{3,g}| \leq G_v/n_v$, we conclude that $|\bar{\eta}_v| = O(G_v/n_v)$. Thus, to complete the proof of (5), it remains to show that $\bar{e}_v = O_p(G_v n_v^{-1})$. Let $f_{gv}(\mathbf{a}) \equiv (e_{gv}(\mathbf{a}))^2$. By the C_r inequality (Sen and Singer 1993, page 21),

$$\begin{aligned} |f_{gv}(\mathbf{a})|^2 &\leq 5^3 (|h(\mathbf{a})|^4 + |h(N_g^{-1} \mathbf{t}_{gv})|^4 \\ &\quad + \sum_{k=1}^3 |h^{(k)}(N_g^{-1} \mathbf{t}_{gv})|^4 |a_k - N_g^{-1} \mathbf{t}_{gv}|^4). \end{aligned}$$

Using (A1) and (A4), straightforward bounding arguments show that $|h(N_g^{-1} \hat{t}_{gv}^*)|^4 = O((n_v/G_v)^4)$ and that $N_g^{-4} |\hat{t}_{k,g} - t_{k,g}|^4 = O(1)$ for $k = 1, 2, 3$. Therefore,

$$|f_{gv}(N_g^{-1} \hat{t}_{gv}^*)|^2 = O\left(\frac{n_v^4}{G_v^4}\right).$$

Since by Lemma 2, $N_g^{-8} E \|\hat{t}_{gv}^* - \mathbf{t}_{gv}\|^8 = O((n_v/G_v)^{-4})$, and $v |f_{gv}(N_g^{-1} \hat{t}_{gv}^*)|^2$ is continuous at any realization of $N_g^{-1} \hat{t}_{gv}^*$, then the sequence $\{|f_{gv}(N_g^{-1} \hat{t}_{gv}^*)|^2\}$ satisfies the conditions of Theorem 5.4.4 (with $\eta = 1$, $p = 4$) of Fuller (1996, page 247). Therefore,

$$E[|f_{gv}(N_g^{-1} \hat{t}_{gv}^*)|^2] = O(1), \quad \forall g = 1, 2, \dots, G_v.$$

Now, from the continuity of $f_{gv}(\cdot)$ and its derivatives up to order three, $\{f_{gv}(N_g^{-1} \hat{t}_{gv}^*)\}$ satisfies the conditions of Theorem 5.4.3 with $\delta = 1$, $s = 4$ and $a_v = O(\sqrt{G_v/n_v})$ of Fuller (1996, pages 244–245). Hence,

$$E f_{gv}(N_g^{-1} \hat{t}_{gv}^*) = O(a_v^4) = O\left(\frac{G_v^2}{n_v^2}\right), \quad \forall g = 1, 2, \dots, G_v,$$

because $f_{gv}(\cdot)$ and all of its derivatives up to order three are zero at $N_g^{-1} \hat{t}_{gv}^*$. Therefore, we conclude that

$$\begin{aligned} E|\bar{e}_v| &\leq \frac{1}{N_v} \sum_{g=1}^{G_v} N_g E|e_{gv}(N_g^{-1} \hat{t}_{gv}^*)| \\ &\leq \frac{1}{N_v} \sum_{g=1}^{G_v} N_g (E f_{gv}(N_g^{-1} \hat{t}_{gv}^*))^{1/2} = O\left(\frac{G_v}{n_v}\right), \end{aligned}$$

which leads to $\bar{e}_v = O_p(G_v n_v^{-1})$ by an application of Markov's inequality.

Expressions (6) and (7) are obtained by direct computation of the moments of the linear estimator \tilde{t}_{WC} under the sampling design and the response mechanism.

Proof of Corollary 3.1: Let

$$Z_v = \frac{1}{V_v^{1/2}} \left(\frac{\tilde{t}_{WC}}{N_v} - \bar{Y}_v - B_v \right)$$

and

$$W_v = \frac{1}{V_v^{1/2}} \left(\frac{\hat{t}_{WC}^*}{N_v} - \frac{\tilde{t}_{WC}}{N_v} \right),$$

where $V_v = \text{Var}(\tilde{t}_{WC}/N_v)$. Hence,

$$\left[\text{Var} \left(\frac{\tilde{t}_{WC}}{N_v} \right) \right]^{1/2} \left(\frac{\hat{t}_{WC}^*}{N_v} - \bar{Y}_v - B \right) = Z_v + W_v.$$

Since $V/n_v V_v \rightarrow 1$, as $v \rightarrow \infty$, then,

$$Z_v = \frac{1}{V^{1/2}} \left(\frac{V}{n_v V_v} \right)^{1/2} n_v^{1/2} \left(\frac{\tilde{t}_{WC}}{N_v} - \bar{Y}_v - B_v \right) \xrightarrow{L} \frac{1}{V^{1/2}} Z,$$

where $Z \sim N(0, V)$. Also, (A7) with $\gamma < 1/2$ implies that $n_v^{1/2} O_p(G_v n_v^{-1}) = o_p(1)$. Hence, by Theorem 3.1,

$$W_v = \frac{1}{V^{1/2}} \left(\frac{V}{n_v V_v} \right)^{1/2} n_v^{1/2} \left(\frac{\hat{t}_{WC}^*}{N_v} - \frac{\tilde{t}_{WC}}{N_v} \right) = o_p(1).$$

The result of the corollary follows, therefore, from Fuller (1996, Theorem 5.2.1).

Proof of Theorem 3.2: Fix a $g \in \{1, 2, \dots, G_v\}$. The conditions of the theorem imply, by the Intermediate Value Theorem, that there exists X_{0g} inside the interval defined by the lowest and the highest values of $X_i \in U_g$ such that $\bar{\varphi}_g = N_g^{-1} \sum_{i \in U_g} \varphi_i = \varphi(X_{0g})$. Also, by the mean Value Theorem, $\forall i \in U_g$,

$$\varphi_i = \varphi(X_i) = \varphi(X_{0g}) + \varphi'(c^*)(X_i - X_{0g}),$$

where c^* is between X_i and X_{0g} . So,

$$|\varphi_i - \bar{\varphi}_g| = |\varphi'(c^*)| |X_i - X_{0g}| \leq C \frac{X_{(N)} - X_{(1)}}{G_v}, \quad (8)$$

for some constant $C \in (0, \infty)$ and, by (A5) and (A6),

$$\left| \text{Bias} \left(\frac{\tilde{t}_{\text{WC}}}{N_v} \right) \right| \leq C \lambda_g^{-1} \lambda_s \frac{X_{(N)} - X_{(1)}}{G_v}.$$

Observe now that since

$$|\tilde{Y}_{ig}| \leq \frac{1}{\pi_i} \frac{\phi_i}{\bar{\phi}_g} |Y_i| + \frac{1}{\pi_i} \frac{|\phi_i - \bar{\phi}_g|}{\bar{\phi}_g} |\tilde{Y}_g|,$$

then, by (A1), (A6) and (8),

$$\tilde{Y}_{ig} = O\left(\frac{N_v}{n_v}\right) + O\left(\frac{N_v}{n_v G_v}\right), \forall U_g, \forall g = 1, 2, \dots, G_v,$$

which implies that

$$\tilde{Y}_{ig} \tilde{Y}_{jg'} = O\left(\frac{N_v^2}{n_v^2}\right) + O\left(\frac{N_v^2}{n_v^2 G_v}\right), \forall U_g, \forall g = 1, 2, \dots, G_v.$$

Using the facts that, by (A7), $N_g/N_v = O(1/G_v)$, by (A2) and (A3), $\sum_{U_g} \sum_{U_{g'}} \Delta_{ij} = O(n_v/G_v)$ and, for $g \neq g'$, $\sum_{U_g} \sum_{U_{g'}} \Delta_{ij} = O(n_v/G_v^2)$, then, the first term of $\text{Var}(\tilde{t}_{\text{WC}}/N_v)$ is bounded by

$$O\left(\frac{1}{n_v}\right) + O\left(\frac{1}{n G_v}\right).$$

Since the second terms of $\text{Var}(\tilde{t}_{\text{WC}}/N_v)$ is bounded by $O(1/n_v)$, the conclusion follows.

REFERENCES

- CASSEL, C.-M., SÄRNDAL, C.-E. and WRETMAN J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete data in sample surveys: Theory and bibliographies*, (Eds. W.G. Madow, I. Olkin and D. B. Rubin). Academic Press, New York: London. 3, 143–160.
- COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 24, 295–313.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3rd Ed.). New York: John Wiley & Sons, Inc.
- DA SILVA, D.N. (2003). Adjustments for Survey Unit Nonresponse Under Nonparametric Response Mechanisms. Ph. D. Thesis, Iowa State University, Ames, IA.
- DA SILVA, D.N., and OPSOMER, J.D. (2003). A kernel smoothing method to adjust for unit nonresponse in sample surveys. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association [CD-ROM]. Alexandria, VA. Article #00605.
- FULLER, W.A. (1996). *Introduction to Statistical Time Series* (Second Edition). Wiley.
- FULLER, W.A., and KIM, J.-K. (2003). Hot deck imputation for the response model. Submitted for publication.
- HANSEN, M.H., MADOW, W.G. and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*. 78, 776–793.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47, 663–685.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*. 77, 89–96.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. Institute of Social Research.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*. 12, 1–16.
- KALTON, G., and MALIGALIG, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. In *Proceedings of the Bureau of the Census Annual Research Conference*. U.S. Bureau of the Census (Suitland, MD). 409–428.
- KIM, J.-K., and FULLER, W.A. (1999). Jackknife variance estimation after hot deck imputation. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA. 825–830.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*. 54, 139–157.
- LITTLE, R.J.A., and RUBIN, D.B. (2002). *Statistical Analysis With Missing Data*. Wiley. 20
- OH, H. L., and SCHEUREN, F.J. (1983). Weighting adjustments for unit non-response. In *Incomplete data in sample surveys: Theory and bibliographies*, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin). Academic Press New York: London. 2, 143–184.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SEN, P.K., and SINGER, J.D.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman and Hall Ltd.
- U.S. BUREAU OF THE CENSUS (1963). The Current Population Survey: A report on methodology. Technical Paper No. 7, Washington, DC.

Variance Estimation with Hot Deck Imputation Using a Model

J. MICHAEL BRICK, GRAHAM KALTON and JAE KWANG KIM¹

ABSTRACT

When imputation is used to assign values for missing items in sample surveys, naïve methods of estimating the variances of survey estimates that treat the imputed values as if they were observed give biased variance estimates. This article addresses the problem of variance estimation for a linear estimator in which missing values are assigned by a single hot deck imputation (a form of imputation that is widely used in practice). We propose estimators of the variance of a linear hot deck imputed estimator using a decomposition of the total variance suggested by Särndal (1992). A conditional approach to variance estimation is developed that is applicable to both weighted and unweighted hot deck imputation. Estimation of the variance of a domain estimator is also examined.

KEY WORDS: Missing data; Model-assisted approach; Conditional variance estimation.

1. INTRODUCTION

The important practical problem of estimating the variance of an estimate computed from a data set in which some of the items are missing and values are assigned by imputation has been addressed in a number of different ways (e.g., see Rubin 1987 and Rao and Shao 1992). The approach used in this article is based on the model-assisted approach introduced by Särndal (1992). In the initial application, Särndal used the model-assisted approach with a simple random sample in which the missing data were imputed using deterministic ratio imputation. Subsequently, the approach has been extended to other imputation methods and sample designs (e.g., Deville and Särndal 1994; Rancourt, Särndal and Lee 1994; and Gagnon, Lee, Rancourt and Särndal 1996). This article extends the model-assisted approach to general forms of linear estimators in which missing values have been assigned by hot deck imputation within imputation cells. This form of hot deck imputation, which replaces a missing item by the value observed for a responding unit in the same cell, is one of the most frequently used methods of imputing for missing items in household sample surveys (Brick and Kalton 1996). This paper employs a conditional approach to develop a variance estimator for hot deck imputed estimators that is valid for general sample designs and a variety of estimation strategies.

In the model-assisted approach, the difference between an imputed estimator (the term used here to denote an estimator based in part on imputed values), $\hat{\theta}_I$, and the corresponding finite population parameter, θ_N , is written as

$$\hat{\theta}_I - \theta_N = (\hat{\theta}_n - \theta_N) + (\hat{\theta}_I - \hat{\theta}_n), \quad (1)$$

where $\hat{\theta}_n$ is the usual, approximately design unbiased, estimator of θ_N with complete response. The first term on the right hand side of (1) is called the sampling error and depends only on the sampling distribution of the estimator based on the sample design used to select the full sample, denoted by p . The second term is the imputation error; it depends on the sampling distribution, the response mechanism (R) that generates the respondents from the full sample, and the imputation mechanism (I) for filling in the missing values. This paper is restricted to estimators $\hat{\theta}_I$ that involve only one variable subject to missing data.

We use a model-assisted approach that makes assumptions about the distribution of the variable of interest in the population. We refer to these assumptions as a superpopulation model, denoted by ξ . In general, the aim of imputation is to create a multi-purpose data set that can be validly analyzed in many different ways, potentially involving the associations of a variable subject to imputation with any of the other variables in the data set. Since a superpopulation model is needed to impute for item non-responses in a way that preserves such associations, it is natural to use that approach also in variance estimation.

Under the superpopulation model, the total variance for an imputed estimator is given by

$$V_{\text{TOT}} = E_{\xi} E_p E_R E_I (\hat{\theta}_I - \theta_N)^2, \quad (2)$$

where E_{ξ} , E_p , E_R , and E_I refer to expectations with respect to the superpopulation model, the sampling mechanism, the response mechanism, and the imputation mechanism, respectively. We assume that the sample design, response mechanism, and the imputation mechanism are unconfounded as described by Rubin (1987) and used by Särndal (1992) and all of the other literature cited above

¹ J. Michael Brick and Graham Kalton, Westat, 1650 Research Blvd., Rockville, MD 20850, U.S.A. E-mail: mikebrick@westat.com; Jae Kwang Kim, Department of Applied Statistics Yonsei University, Seoul 120-749, Korea.

on the model-assisted approach. Essentially, unconfounded mechanisms allow the order of the expectations to be changed so that the expectation with respect to the model can be taken first. Thus, the total variance can be re-written as $V_{\text{TOT}} = E_p E_R E_I E_\xi (\hat{\theta}_I - \theta_N)^2$. Roughly speaking, unconfounded sampling, response, and imputation mechanisms imply that the mechanisms are independent of the distribution of the y -value being analyzed after conditioning on auxiliary variables (e.g., stratification variables for sampling or imputation cells for imputing). Thus, for example, we assume the value of the variable being imputed is independent of the probability of response within each hot-deck cell. Rubin (1987, pages 36-39) has a more detailed discussion of unconfounded mechanisms.

Using the decomposition given in equation (1), Särndal (1992) expressed the total variance for the imputed estimator as

$$V_{\text{TOT}} = E_\xi E_p E_R E_I (\hat{\theta}_I - \theta_N)^2 = V_{\text{SAM}} + V_{\text{IMP}} + 2V_{\text{MIX}}, \quad (3)$$

where $V_{\text{SAM}} = E_\xi E_p E_R \hat{(\theta}_n - \theta_N)^2$ is the sampling variance, $V_{\text{IMP}} = E_\xi E_p E_R E_I (\hat{\theta}_I - \hat{\theta}_n)^2$ is the imputation variance, and $V_{\text{MIX}} = E_\xi E_p E_R E_I [(\hat{\theta}_I - \hat{\theta}_n)(\hat{\theta}_n - \theta_N)]$ is a mixed component. In this formulation, the total variance and its components are more aptly described as anticipated variances because they incorporate the added expectation with respect to the superpopulation model.

The model-assisted approach to variance estimation with imputed data used in this paper should be distinguished from model-assisted sampling (Särndal, Swenson and Wretman 1992). With model-assisted sampling, models are used to guide the choice of efficient sample designs and estimators, but the validity of statistical inferences is not dependent on the validity of the models. In contrast, when some data are missing, reliance on models for inferences is essential, both for point estimators and for variance estimators for them. In this paper, the general approach to inference employs the imputation model assumptions (i.e., superpopulation model and unconfoundedness assumptions) only to the extent necessary to account for imputed data. Both the point estimators and the variance estimators are the standard design-based estimators when no data are missing. Whether the variance estimators are approximately unbiased for V_{SAM} depends on the validity of the imputation model. Also, the estimators for V_{IMP} and V_{MIX} rely completely on the imputation model. Thus the validity of the model is much more critical with model-assisted variance estimation with imputed data than it is with model-assisted sampling. Särndal (1992) argues that if we are willing to accept the validity of the model in point estimation with imputed data, we should also be willing to accept its validity for variance estimation.

Variance estimators are obtained by conditioning on the realized set of sampled units, responding units, and imputations. We develop estimators of $V_{\text{SAM}} = E_\xi [(\hat{\theta}_n - \theta_N)^2 | \mathbf{A}, \mathbf{A}_R, \mathbf{d}]$, $V'_{\text{IMP}} = E_\xi [(\hat{\theta}_I - \hat{\theta}_n)^2 | \mathbf{A}, \mathbf{A}_R, \mathbf{d}]$, and $V'_{\text{MIX}} = E_\xi [(\hat{\theta}_n - \theta_N)(\hat{\theta}_I - \hat{\theta}_n) | \mathbf{A}, \mathbf{A}_R, \mathbf{d}]$, where \mathbf{A} and \mathbf{A}_R denote matrices of indices for the sampled and responding units, respectively, and \mathbf{d} is the set of indices for the imputations. The conditioning is on the set of indices, not on the values of the units. The matrix \mathbf{d} is an $r \times (n - r)$ matrix in which the rows refer to respondents and the columns to nonrespondents. In this paper, we consider only single imputation methods, in which case all but one of the $d_{ij} = 0$ in every column. The exception occurs in the row of the donor respondent when $d_{ij} = 1$.

By considering the conditional expectations of V'_{IMP} and V'_{MIX} , the estimators reflect the number of times responding units are used as donors in the given application rather than taking the expectation over all possible imputation outcomes. We argue below that these are the appropriate variances to estimate in a given application. If the variance estimators are conditionally unbiased, they are also, of course, unconditionally unbiased.

A conditional approach is useful for two reasons. First, when an estimator is conditionally unbiased and consistent (as $\hat{\theta}_I$ is assumed to be for $\hat{\theta}_n$), the conditional variance is generally a more appropriate estimator for making inferences from a realized sample than an unconditional variance (Holt and Smith 1979, Rao 1999, Kalton 2002). Thus, a variance estimator that conditions on the actual number of times each donor is used is to be preferred to a variance estimator that averages over all possible donor selections. Second, the results apply to any unconfounded sampling, response, and imputation mechanisms that produce the same set of sampled units, respondents, and imputations. Therefore, the results given below for hot deck imputation apply to any unconfounded imputation scheme that substitutes observed values for missing ones and for which $E_\xi(\hat{\theta}_I) = E_\xi(\hat{\theta}_n)$.

2. HOT DECK IMPUTATION

We consider a simple model for which hot deck imputation is appropriate. Assume that the finite population (U) is composed of G classes or cells. Within cell g ($g = 1, \dots, G$), the elements in U are realizations of independently and identically distributed random variables with mean μ_g and variance σ_g^2 . This cell mean model can be written as

$$Y_i^{i,i} | \mu_g, \sigma_g^2, i \in U_g. \quad (4)$$

where $i.i.$ is an abbreviation for independently and identically distributed.

A linear estimator of θ_N with complete item response from a complex sample survey can be written as

$$\hat{\theta}_n = \sum_{i \in A} w_i y_i, \quad (5)$$

where w_i is the weight that accounts for unequal selection probabilities and the estimation strategy. When the cell mean model holds, a more efficient estimator of θ_N uses the unweighted group means, *i.e.*, $\hat{\theta}_n' = \sum \sum w_{gi} \bar{y}_g$ where $\bar{y}_g = \sum_i y_{gi} / n_g$. However, the model-assisted approach does not place complete reliance on the model; rather, it uses the standard design-based approach to the extent possible and the model is used only for the missing data. The weights in (5) can be the inverse of the probability of selection weights or calibration adjusted weights, as described below.

The hot deck imputed value for y_i is $y_j^* = \sum_{i \in A_R} d_{ij} y_i$ and the imputed estimator is

$$\hat{\theta}_I = \sum_{i \in A} w_i \tilde{y}_i = \sum_{i \in A_R} w_i y_i + \sum_{j \in A_M} w_j \sum_{i \in A_R} d_{ij} y_i, \quad (6)$$

where $\tilde{y}_i = y_i$ for $i \in A_R$ and $\tilde{y}_i = y_i^*$ for $i \in A_M$. We assume throughout that imputed values are selected from respondents in the same imputation cells, and that each cell contains at least one respondent.

This imputation formulation does not specify the way in which donors are selected. It thus covers both unweighted hot deck imputation in which donors are selected with equal probabilities within each cell and weighted hot deck imputation. Weighted hot decks are typically used when assumptions are made only about the response distribution. The form (6) also covers with and without replacement imputation methods. For example, it covers the common hot deck procedure in which a respondent is randomly selected to be a donor within a cell, and then that respondent is not used as a donor again until every other respondent in the cell has been used.

While not explicitly considered here, nearest neighbor imputation procedures that use continuous variables to identify a small set of the most similar respondents and then randomly select one as the donor, satisfy the above requirements. Furthermore, researchers often use hot deck methods even when continuous variables are available. Little (1986) discusses strategies for forming imputation cells using variables that are predictive of the y -variable and notes that imputation within cells and regression imputation should produce similar results in many circumstances. Cochran (1968) and Aigner, Goldberger and Kalton (1975) show that a relatively small number of well-constructed cells

formed from a continuous variable can capture a large proportion of the predictive power of the variable.

The conditional bias of the imputed estimator under the cell mean model is

$$E_{\xi}(\hat{\theta}_I - \hat{\theta}_n | \mathbf{A}, \mathbf{A}_R, \mathbf{d}) = E_{\xi} \left[\sum_{j \in A_M} w_j (y_j^* - y_j) | \mathbf{A}, \mathbf{A}_R, \mathbf{d} \right] = 0,$$

since $E_{\xi}(y_j^*) = E_{\xi}(\sum_{i \in A_R} d_{ij} y_i) = \sum_{i \in A_R} d_{ij} E_{\xi}(y_i) = \sum_{i \in A_R} d_{ij} \mu_g = \mu_g$ for j in cell g . This expectation is conditioned on the indices of the sampled units, the responding units, and the donors. However, since the estimator is conditionally unbiased for any sample, it is also unconditionally unbiased. Kim and Fuller (1999) also use this conditioning argument. Estimators for each component of the variance of the hot deck imputed estimator are given in the next section.

3. ESTIMATION OF THE COMPONENTS OF THE TOTAL VARIANCE

This section contains the main results about estimators of the three components of the total variance of a linear hot deck imputed estimator. Throughout, we assume unconditional sampling, response, and imputation mechanisms and a linear complete sample estimator of the form (5). The results require that the cell mean model holds and that there is at least one respondent in each imputation cell. We begin with the variance due to sampling, V_{SAM} .

We assume that there exists a complete sample variance estimator, \hat{V}_n , that is design unbiased for the sampling variance of $\hat{\theta}_n$, is a quadratic in the y -variable, and is of the form

$$\hat{V}_n = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} y_i y_j = \sum_{i \in A} \sum_{i,j \in A} \Omega_{ii} y_i^2 + 2 \sum_{i < j} \Omega_{ij} y_i y_j, \quad (7)$$

for known coefficients Ω_{ij} . This formulation covers the Horvitz-Thompson estimator, where the Ω_{ij} are determined by the single and joint probabilities of selection. It also covers the linearized variance estimator for the generalized regression (GREG) estimator. Rao, Yung and Hidirolglou (2002) show that the linearized variance estimator for the GREG estimator can be written by substituting $g_{is} e_i$ for y_i in the variance estimator for the Horvitz-Thompson estimator of a total. Here, g_{is} is the sample-dependent g -weight and $e_i = y_i - \mathbf{x}_i' \hat{\mathbf{B}}$, where \mathbf{x}_i is the vector of auxiliary variables and $\hat{\mathbf{B}}$ is the vector of estimated regression coefficients. Since g_{is} is not a function of y and $\hat{\mathbf{B}}$ is linear in the y -variable, $g_{is} e_i$ is linear in y . Therefore, the linearized variance estimator for the GREG estimator is quadratic in y and can be expressed in the form given by

equation (7). Note that in this case the Ω_{ij} may be dependent on the specific sample as well as on the selection probabilities. Deville and Särndal (1992) show that any calibration estimator has the same asymptotic variance as the GREG. Thus, asymptotic variance estimators for calibration estimators in general have the required quadratic form.

The naïve variance estimator treats imputed values as if they were observed values and can be written as

$$\hat{V}_0 = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} \tilde{y}_i \tilde{y}_j. \quad (8)$$

Lemma 1 gives the bias of the naïve variance estimator as an estimator for V'_{SAM} . As noted earlier, the naïve variance estimator is proposed as the estimator of V'_{SAM} to be as consistent as possible with design-based inference. An additional practical reason for using the naïve variance estimator is to take advantage of existing software programs that estimate the sampling variance under complex samples.

Lemma 1. *Under the cell mean model with unconfounded sampling, response, and imputation mechanisms and the assumptions that $\hat{\theta}_1$ is an unbiased hot deck imputed linear estimator given by (6) and \hat{V}_n is an unbiased complete sample variance estimator given by (7), then the bias of the naïve variance estimator, \hat{V}_0 , as an estimator of V_n is*

$$2 \sum_{g=1}^G \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} \Omega_{ij} d_{ij} \sigma_g^2 + 2 \sum_{g=1}^G \sum_{i < j} \sum_{i, j \in A_{M_g}} \Omega_{ij} \gamma_{ij} \sigma_g^2, \quad (9)$$

where $A_{R_g} = A_R \cap U_g$, $A_{M_g} = A_M \cap U_g$, and

$$\gamma_{ij} = \sum_{k \in A_R} d_{ki} d_{kj}. \quad (10)$$

For any two nonrespondents, i and j , that have the same donor, $\gamma_{ij} = 1$; $\gamma_{ij} = 0$ otherwise. By definition, $\gamma_{ii} = 1$.

Proof. We begin by noting that the difference between \hat{V}_0 and \hat{V}_n can be written as:

$$\begin{aligned} \hat{V}_0 - \hat{V}_n &= \sum_{i \in A} \Omega_{ii} (\tilde{y}_i^2 - y_i^2) \\ &\quad + 2 \sum_{i < j} \sum_{i, j \in A} \Omega_{ij} (\tilde{y}_i \tilde{y}_j - y_i y_j) \\ &= \sum_{i \in A_M} \Omega_{ii} (y_i^{*2} - y_i^2) \\ &\quad + 2 \sum_{i < j} \sum_{i \in A_R, j \in A_M} \Omega_{ij} (y_i^* y_j^* - y_i y_j) \\ &\quad + 2 \sum_{i < j} \sum_{i, j \in A_M} \Omega_{ij} (y_i^* y_j^* - y_i y_j). \end{aligned} \quad (11)$$

Under the cell mean model, the conditional expectation of the first term of (11) is zero. The conditional expectation $E_\xi(y_i y_j^* - y_i y_j) = E_\xi[y_i (y_j^* - y_j)] = 0$ unless respondent i is the donor for nonrespondent j ; it is thus zero when units i and j are in different cells and is only nonzero for one i and j in the same cell g . It may be represented by $E_\xi[y_i (y_j^* - y_j)] = d_{ij} \sigma_g^2$. The conditional expectation $E_\xi(y_i^* y_j^* - y_i y_j)$ is zero unless nonresponding units i and j have the same donor, which can occur only if these units are in the same cell. It can be represented by $E_\xi(y_i^* y_j^* - y_i y_j) = \gamma_{ij} \sigma_g^2$ for $i \neq j$. Applying these results in equation (11) gives

$$\begin{aligned} E_\xi(\hat{V}_0 - \hat{V}_n | \mathbf{A}, \mathbf{A}_R, \mathbf{d}) &= 2 \sum_{g=1}^G \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} \Omega_{ij} d_{ij} \sigma_g^2 \\ &\quad + 2 \sum_{g=1}^G \sum_{i < j} \sum_{i, j \in A_{M_g}} \Omega_{ij} \gamma_{ij} \sigma_g^2. \end{aligned} \quad (12)$$

The proof is completed by noting that since \hat{V}_n is unbiased under the design, it is also unbiased for V'_{SAM} . Substituting a model unbiased estimator for σ_g^2 , say $\hat{\sigma}_g^2$, gives an unbiased estimator of the bias of the naïve variance estimator. Note that whenever respondents donate their values to more than one nonrespondent, the last term in equation (12) is positive; otherwise, it is zero.

Two simple examples illustrate applications of these results. Consider first the estimation of a population mean from a simple random sample selected with replacement. In this case, $\Omega_{ii} = n^{-2}$ and $\Omega_{ij} = -n^{-2}(n-1)^{-1}$ for $i \neq j$. Assume that the cell mean model holds with hot deck imputation and that no donor is used more than once. By Lemma 1, the bias of \hat{V}_0 is $-2n^{-2}(n-1)^{-1} \sum_g m_g \sigma_g^2$, where $m_g = \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} d_{ij}$ is the number of imputed values in cell g . In this case, the bias of the naïve variance estimator is $O_p(n^{-2})$ and hence is negligible for large n . Now suppose that every missing value in each cell is imputed from the same donor. In this case, with $\sum_{i < j} \gamma_{ij} = m_g(m_g - 1)/2$, the bias of \hat{V}_0 is $-n^{-2}(n-1)^{-1} \sum_g (m_g^2 + m_g) \sigma_g^2$, which is $O_p(n^{-1})$ and is the same order as the sampling variance.

As the second example, consider a simple two-stage sample of size $n = ab$, in which a clusters are selected from a population of A equal-sized clusters by simple random sampling and b of B elements are selected by simple random sampling within each sampled cluster. Let y_{ai} be the value for y for sampled unit i in cluster a . Assume that the first stage sampling fraction is small enough to ignore. The estimate of the variance of the sample mean is of the form given by equation (7) where $\Omega_{ai, \beta j} = a^{-2} b^{-2} = n^{-2}$ for $\alpha = \beta$, and $\Omega_{ai, \beta j} = -n^{-2}(a-1)^{-1}$ for $\alpha \neq \beta$. These

values can now be inserted into equation (9) to compute an estimate of the bias. For example, suppose that all missing values are imputed using donors from the same cluster (the cells are the clusters) and no donor is used more than once. In this case, the bias of the naïve variance estimator is $2n^{-2} \sum_a m_a \sigma_a^2$, where m_a is the number of nonrespondents in cluster a . Now, suppose an overall cell mean model hot deck is used and no donor can donate more than once, but that donors are always chosen from different clusters than their missing values. In this case, the bias of the naïve variance estimator is $-2n^{-2} (a-1)^{-1} \sigma^2 \sum_a m_a$. This two-stage example shows the naïve variance estimator can be biased in either direction. In both of the cases considered, the bias is of lower order than the variance, and if a is large the bias will be negligible.

The second component of the total variance is the variance due to imputation, V_{IMP} . Lemma 2 gives an unbiased estimator for this component with hot deck imputation.

Lemma 2. *Under the assumptions used in Lemma 1, an unbiased estimator of V'_{IMP} is*

$$\hat{V}'_{\text{IMP}} = 2 \sum_{g=1}^G \left\{ \sum_{i \in A_{M_g}} w_i^2 \hat{\sigma}_g^2 + \sum_{i < j} \sum_{i, j \in A_{M_g}} w_i w_j \gamma_{ij} \hat{\sigma}_g^2 \right\}. \quad (13)$$

where $\hat{\sigma}_g^2$ is an unbiased estimator for σ_g^2 .

Proof. Since the variance due to imputation involves the squared difference between the imputed and complete response estimates, we begin by writing

$$\begin{aligned} (\hat{\theta}_I - \hat{\theta}_n)^2 &= \left[\sum_{i \in A} w_i (\tilde{y}_i - y_i) \right]^2 \\ &= \sum_{i \in A_M} w_i^2 (y_i^* - y_i)^2 \\ &\quad + 2 \sum_{i < j} \sum_{i, j \in A_M} w_i w_j (y_i^* - y_i)(y_j^* - y_j). \end{aligned}$$

Noting that $E_{\xi}(y_i^* - y_i)^2 = 2\sigma_g^2$ for i in cell g and, from above, $E_{\xi}[(y_i^* - y_i)(y_j^* - y_j)] = E_{\xi}(y_i^* y_j^* - y_i y_j) = \gamma_{ij} \sigma_g^2$, it follows that

$$V'_{\text{IMP}} = 2 \sum_{g=1}^G \left\{ \sum_{i \in A_{M_g}} w_i^2 \sigma_g^2 + \sum_{i < j} \sum_{i, j \in A_{M_g}} w_i w_j \gamma_{ij} \sigma_g^2 \right\}. \quad (14)$$

Substituting $\hat{\sigma}_g^2$, a model unbiased estimator for σ_g^2 , establishes the lemma.

Equation (14) shows that the imputation variance has positive contributions from each imputed value and also from using donors more than once. For example, suppose that the weights for all sampled cases are equal. The

contribution to the imputation variance from cell g is then proportional to the sum of the number of missing cases in the cell and the number of pairs of nonrespondents that receive values from the same donors. Limiting the number of times donors are re-used can reduce the imputation variance.

The third term in the total variance is V_{MIX} , which previous research often considered small or negligible (e.g., Särndal 1992; Deville and Särndal 1994). Lemma 3 gives an unbiased estimator for V'_{MIX} .

Lemma 3. *Under the assumptions used in Lemma 1, an unbiased estimator for V'_{MIX} is*

$$\sum_{g=1}^G \left[\sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} w_i w_j d_{ij} - \sum_{j \in A_{M_g}} w_j^2 \right] \hat{\sigma}_g^2. \quad (15)$$

Proof. Begin by writing $(\hat{\theta}_I - \hat{\theta}_n)(\hat{\theta}_n - \theta_N) = \hat{\theta}_n(\hat{\theta}_I - \hat{\theta}_n) - \theta_N(\hat{\theta}_I - \hat{\theta}_n)$. Let θ_N be the finite population total, which can be written as $\sum_{i \in U-A} y_i + \sum_{i \in A_R} y_i + \sum_{i \in A_M} y_i$. Using this expression, the second component can be expanded as

$$\begin{aligned} \theta_N(\hat{\theta}_I - \hat{\theta}_n) &= \\ &\left(\sum_{i \in U-A} y_i + \sum_{i \in A_R} y_i + \sum_{i \in A_M} y_i \right) \left[\sum_{j \in A_M} w_j (y_j^* - y_j) \right]. \end{aligned}$$

In taking the conditional expectation of this product, the only nonzero contributions occur either when unit i in A_R is the donor for y_j^* , or when unit i in A_M in the first set of parentheses is unit j in the second set. In the first case, $E_{\xi}[y_i(y_j^* - y_j)] = d_{ij} \sigma_g^2$ for $i \in A_{R_g}$, $j \in A_{M_g}$. In the second case, if nonrespondent unit i in A_{M_g} is the same as unit j in the second term, $i = j$, $E_{\xi}[y_i(y_j^* - y_j)] = -\sigma_g^2$, and this expectation is 0 otherwise. Thus,

$$\begin{aligned} E_{\xi}(\theta_N(\hat{\theta}_I - \hat{\theta}_n) | \mathbf{A}, \mathbf{A}_R, \mathbf{d}) &= \sum_g \sum_{j \in A_{M_g}} w_j \sigma_g^2 \\ &\quad - \sum_g \sum_{j \in A_{M_g}} w_j \sigma_g^2 = 0. \end{aligned}$$

The first term can be expressed as

$$\begin{aligned} \hat{\theta}_n(\hat{\theta}_I - \hat{\theta}_n) &= \\ &\left(\sum_{i \in A_R} w_i y_i + \sum_{i \in A_M} w_i y_i \right) \left[\sum_{j \in A_M} w_j (y_j^* - y_j) \right]. \end{aligned}$$

Using the results for $E_{\xi}(y_i(y_j^* - y_j))$ given above,

$$V'_{\text{MIX}} = \sum_{g=1}^G \left(\sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} w_i w_j d_{ij} - \sum_{g=1}^G \sum_{i \in A_{M_g}} w_i^2 \right) \sigma_g^2. \quad (16)$$

Substituting an unbiased estimator of σ_g^2 proves the lemma.

The estimator of V'_{MIX} is zero when the weights are constant, or more generally when the weights of the donors are equal to the weights of the missing cases to which they are assigned. Most of the simulations in the literature (e.g., Särndal 1992; Lee, Rancourt and Särndal 1995) have used simple random samples so that the estimates of the mixed term from the simulations are approximately equal to zero.

To illustrate the effect of unequal weights, consider a stratified simple random sample selected from two equal size strata with replacement, and suppose that the sampling rate in stratum 2 is k times the rate in stratum 1. Let the imputation model be the overall cell mean model and let the hot deck procedure select donors with simple random sampling without replacement. For this simple situation, V'_{MIX} can be derived algebraically. Table 1 shows the percentage contribution of the mixed term to the total variance ($100 \cdot 2V'_{\text{MIX}}/V'_{\text{TOT}}$) for various combinations of strata response rates. The table illustrates the fact that when the sampling weights are unequal, the contribution of the mixed term may be important and can be either positive or negative. The mixed term may also be important in domain estimation, as discussed in the next section.

Table 1

Percentage Contribution of the Mixed Term to V'_{TOT}

Response rate		Oversampling rate in stratum 2		
Stratum 1	Stratum 2	$k = 2$	$k = 4$	$k = 6$
100%	80%	4.3	5	13.7
100	60	8.7	10.8	18.3
100	40	13.7	18.3	17.7
100	20	19.9	28.8	29.7
60	100	-15.4	-34.1	-44.5
60	80	-10.4	-27.1	-37.6
60	60	-5.2	-19	-29.3
60	40	1	-8.8	-18.2
60	20	9.4	6.5	0

Now consider estimating the total variance using the three lemmas for the hot deck estimator under the cell mean model. To estimate V'_{SAM} we can either use the naïve variance estimator, with its bias as given in Lemma 1, or correct for the bias with a procedure similar to that recommended by Särndal (1992). For a single stage sample, the bias correction given by Lemma 1 is easy to apply. However, with multi-stage sampling the correction involving Ω may be complicated and difficult to implement in practice. In this case, the naïve variance estimator should produce an adequate approximation provided that the number of sampled clusters is large, that no donor is used too often, and that the percentage of missing data in each cell is not extremely large.

For the other two components, the only unknown quantities that must be estimated from the sample are the cell variances, $\hat{\sigma}_g^2$. These parameters could be estimated using either unweighted observations or weighted observations, where the weights are the selection weights. Fuller (2002) recommends the use of weighted observations to provide more robust estimates. Unbiased estimators of the conditional variance due to imputation and the mixed component are computed by substituting unbiased estimates of the cell variances, $\hat{\sigma}_g^2$. Then, adding \hat{V}_0 , \hat{V}'_{IMP} , and $2\hat{V}'_{\text{MIX}}$ gives an estimator of the total variance

$$\hat{V}'_{\text{TOT}} = \hat{V}_0 + 2 \sum_{g=1}^G \sum_{i < j} \sum_{i, j \in A_{M_g}} w_i w_j \gamma_{ij} \hat{\sigma}_g^2 + 2 \sum_{g=1}^G \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} w_i w_j d_{ij} \hat{\sigma}_g^2. \quad (17)$$

To examine this estimator, we give a few simple examples with known solutions. All of these examples involve samples with equal weights so the mixed component is zero. First, assume simple random sampling with replacement, hot deck imputation under the overall cell mean model, and no donor used more than once. Using the naïve variance estimator for V'_{SAM} , the estimated total variance is $n^{-1} s_y^2 + 2n^{-1} \hat{\sigma}^2 (1 - m^{-1})$, where $s_y^2 = (n-1)^{-1} \sum_{i \in A} (\tilde{y}_i - \bar{y}_l)^2$, r is the number of respondents, and m is the number of missing cases. If we use $\hat{\sigma}^2$ instead of s_y^2 (where $\hat{\sigma}^2$ is model unbiased while s_y^2 has a small sample bias), then this simplifies to $r^{-1} \hat{\sigma}^2 [1 + m(r-m)n^{-2}]$. Taking the expectation of this estimator gives the unconditional variance of the without-replacement hot deck estimator given by Kalton (1983, page 25, 2.3.1.7).

If a multiple cell mean model rather than an overall cell mean model is used, then the estimated total variance is $n^{-1} s_y^2 + 2n^{-2} \sum_{g=1}^G \hat{\sigma}_g^2 (n_g - r_g)$, which is similar to the result given by Tollefson and Fuller (1992).

Continuing with the simple random sampling example, now allow donors to be used more than once with the overall cell mean model. Again using $\hat{\sigma}^2$ instead of s_y^2 , the estimated total variance is approximately

$$n^{-2} \hat{\sigma}^2 \left(n + m + \sum_{i < j} \sum_{i, j \in A_M} \gamma_{ij} \right). \quad (18)$$

For fixed m , the variance in equation (18) is minimized when no donor is used more often than any other donor, to the extent possible (thereby minimizing $\sum_{i \in A_M} \sum_{j \in A_M} \gamma_{ij}$). Therefore, an imputation scheme that uses any donor at most once more than any other donor minimizes the total variance.

If donors are selected by simple random sampling with replacement, then $E_I[\gamma_{ij}] = r^{-1}$ and the expected value of (18) is $r^{-1} \delta^2 [1 + n^{-2} m(r-1)]$. This is the expected variance of the with-replacement hot deck estimator given by Kalton (1983, page 26, 2.3.1.9).

These examples show that the approach produces reasonable estimates for the total variance in simple cases and highlights the conditional nature of the variance estimates. For example, (18) is conditional on the actual number of times donors are used rather than on the expected number of times they are used (the unconditional result). The approach is flexible enough to allow a variety of imputation methods, including with- and without-replacement and weighted and unweighted versions of the hot deck.

4. DOMAIN ESTIMATION

This section considers the important problem of domain estimation under the cell mean model with hot deck imputed data. Previous research on this topic is limited (Lee *et al.* 1995). The standard estimator with complete response for a population total for domain v is $\hat{\theta}_{n_v} = \sum_{i \in A_v} w_i' y_i$, which may be alternatively expressed as $\hat{\theta}_{n_v} = \sum_{i \in A} w_i' y_i$ where $w_i' = \delta_{vi} w_i$ with $\delta_{vi} = 1$ if $i \in A_v$ and $\delta_{vi} = 0$ otherwise. The hot deck imputed estimator is $\hat{\theta}_{v_i} = \sum_{i \in A} w_i \delta_{vi} \tilde{y}_i = \sum_{i \in A} w_i' \tilde{y}_i$. Throughout we assume that δ_{vi} is known for all $i \in A$.

The cell mean model assumes that all the elements in a cell have the same distribution. In general, some elements in a cell may be in the domain and others not. One version of the model assumes a separate cell mean model for the domain alone and then applies an appropriate imputation scheme. The theory given in the previous section covers this case, and it will, therefore, not be discussed further here. While it is feasible to account for key domains in the imputation stage, it is impossible to consider all possible domains analysts may wish to study. Thus, the focus in this section on domains that cut across imputation cells has important practical implications, especially for analysis of public use data files.

We now discuss the estimation of the three components of V'_{TOT} , the variance of an imputed domain total. Consider first the estimation of V'_{SAM_v} . In the case of complete response, by setting $y_i = 0$ for elements outside the domain, the estimated sampling variance can be expressed in the form of equation (7) as $\hat{V}_{n_v} = \sum_{i \in A_v} \Omega_{ii} y_i^2 + 2 \sum \sum_{i < j \in A_v} \Omega_{ij} y_i y_j$. With domain membership known for all sample elements, the conditional bias of the imputed variance estimator \hat{V}_{0_v} , following the developments in section 3 is:

$$E_{\xi} \left(\hat{V}_{0_v} - \hat{V}_{n_v} \mid \mathbf{A}, \mathbf{A}_R, \mathbf{d} \right) = 2 \sum_{g=1}^G \sum_{i \in A_{R_{gv}}} \sum_{j \in A_{M_{gv}}} \Omega_{ij} d_{ij} \sigma_g^2 + 2 \sum_{g=1}^G \sum_{i < j} \sum_{i, j \in A_{M_{gv}}} \Omega_{ij} \gamma_{ij} \sigma_g^2 \quad (19)$$

As discussed in section 3, with large samples \hat{V}_{0_v} may be conveniently employed to estimate V'_{SAM_v} using standard survey sampling variance estimation software. It is interesting to note that the naïve variance estimator would be unbiased if all the donors were from outside the domain (thus, $d_{ij} = 0$) and no donor was used more than once ($\gamma_{ij} = 0$).

The derivation of \hat{V}'_{IMP_v} follows directly from Lemma 2, where the weights are treated as constants in the conditional expectation. Replacing w_i' for w_i in equation (13) gives

$$\begin{aligned} \hat{V}'_{IMP_v} &= 2 \sum_{g=1}^G \left\{ \sum_{i \in A_{M_g}} w_i'^2 \hat{\sigma}_g^2 + \sum_{i < j} w_i' w_j' \gamma_{ij} \hat{\sigma}_g^2 \right\} \\ &= 2 \sum_{g=1}^G \left\{ \sum_{i \in A_{M_{gv}}} w_i'^2 \hat{\sigma}_g^2 + \sum_{i < j} w_i' w_j' \gamma_{ij} \hat{\sigma}_g^2 \right\}. \end{aligned}$$

\hat{V}'_{IMP} does not depend on whether donors come from within or from outside the domain.

The derivation of \hat{V}'_{MIX_v} also follows from section 3. Substituting w_i' for w_i in equation (15) gives

$$\begin{aligned} \hat{V}'_{MIX_v} &= \sum_{g=1}^G \left(\sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} w_i' w_j' d_{ij} - \sum_{j \in A_{M_g}} w_j'^2 \right) \hat{\sigma}_g^2 \\ &= \sum_{g=1}^G \left(\sum_{i \in A_{R_{gv}}} \sum_{j \in A_{M_{gv}}} w_i' w_j' d_{ij} - \sum_{j \in A_{M_{gv}}} w_j'^2 \right) \hat{\sigma}_g^2. \quad (20) \end{aligned}$$

Note that the mixed component is not zero for a domain total, even if all the original weights are equal. With equal weights w (but not equal w'), the contribution to \hat{V}'_{MIX} is zero when the donor is from inside the domain whereas it is negative when the donor is from outside the domain. As a result, $\hat{V}'_{MIX_v} = -w^2 \sum_g l_{gv} \hat{\sigma}_g^2$, where l_{gv} is the number of donors from outside the domain in cell g . In this case, ignoring the mixed component with domain estimation results in an overestimate of the total variance. With unequal weights, the bias due to ignoring the mixed component can be either positive or negative.

The total variance of a (linear) imputed domain estimator under the cell mean model is then estimated by

$$\begin{aligned}\hat{V}'_{TOT_v} = & \hat{V}'_{0_v} + 2 \sum_{g=1}^G \sum_{\substack{i < j \\ i, j \in A_{M_g}}} w_i' w_j' \gamma_{ij} \hat{\sigma}_g^2 \\ & + 2 \sum_g \sum_{i \in A_{R_g}} \sum_{j \in A_{M_g}} w_i' w_j' d_{ij} \hat{\sigma}_g^2.\end{aligned}\quad (21)$$

As an illustration, consider the case of equal weights within the domain ($w_{iv} = w_v$) and no donor used more than once. In this case, the second term on the right in (21) is zero and the third term reflects the variance increase from imputation. If all the missing values are imputed using donors from the domain, then the third term is $2w_v^2 \sum m_{gv} \hat{\sigma}_g^2$ where m_{gv} is the number of missing items in cell g and domain v . On the other hand, if no units are imputed from within the domain, then this term is zero. Thus, the total variance is minimized when the donors are selected from outside the domain rather than from within the domain. This result occurs because imputing from outside the domain in effect substitutes a new value for a missing value for domain estimation, thus maintaining the original domain sample size. On the other hand, imputing from within the domain does not increase domain sample size and there is also a penalty to the variance from reusing a domain respondent's value for the nonrespondent.

If the distribution of y varies by domain (*i.e.*, the imputation model is misspecified), then choosing donors from outside the domain results in biased estimates. Since all models are misspecified to some degree, it is therefore generally unwise to intentionally select donors from outside the domain in order to minimize the variance.

5. SIMULATION STUDY

A small simulation study was performed to examine the model-assisted variance estimates for estimating an overall total and a domain total. A sample of 40 clusters with exactly 5 units in each cluster was selected from an infinite superpopulation, where y_{ai} is the study variable for unit i in cluster α . The y -values were generated from $y_{ai} = \tau a_\alpha + e_{ai}$, where a_α and e_{ai} are independent random draws from the standard normal distribution. Thus, the y -values have mean zero, variance ($\tau^2 + 1$), and correlation $\rho = \tau^2 / (1 + \tau^2)$ if the units are from the same cluster and $\rho = 0$ otherwise. Values of $\tau = 0$ and $\tau = 0.5$ were chosen, giving correlations of 0 and 0.2, respectively. The value, $\rho = 0.2$, was chosen to illustrate the effect of a high intraclass correlation. In addition to the y -variable, an indicator variable for domain v was generated by independent sampling with the probability of being in the domain of 0.25. Respondents were selected from the full sample using a uniform response

probability of 0.6 and missing values were imputed using a single-cell with-replacement hot deck. A total of 5,000 Monte Carlo samples was selected.

The simulated point estimators for the overall total and the domain total are unbiased. The means and biases of the model-assisted variance estimators (\hat{V}'_{TOT}) are given in Table 2 (the tabulated values are divided by $N^2 10^{-4}$). When $\rho = 0$, the relative biases of the variance estimators for the overall and domain totals are very small. On the other hand, when $\rho = 0.2$, the variance estimators have negative relative biases that are not negligible (a relative bias of -13% for the overall total and -5% for the domain total). To identify the source of the bias, Table 2 also gives the means and biases of the three variance components. The tabled values show that V'_{IMP} and V'_{MIX} are approximately unbiased, and it is only \hat{V}'_0 that has a non-negligible bias.

When $\rho = 0$ the cell mean model holds and \hat{V}'_0 is unbiased as expected under the theory. When $\rho = 0.2$, the correlation of the y -values within clusters implies that the cell mean model assumption does not hold. The imputation procedure replaces some missing values using donors from outside the cluster, causing \hat{V}'_0 to underestimate the sampling variance due to the underestimation of the intraclass correlation. In this particular situation, the model failures do not result in biased estimates for the other two components. However, these components could be biased under other types of model failure. The simulation illustrates the dependence of the model-assisted estimators on the model assumptions and this is discussed further in the next section.

Table 2

Mean and Bias of Simulated Variance Estimators, with Cluster Sampling of 40 Clusters with 5 Elements and Response Rate of 60 Percent^a

Estimate	ρ	\hat{V}'_{TOT}		\hat{V}'_0		\hat{V}'_{IMP}		\hat{V}'_{MIX}	
		Mean	Bias	Mean	Bias	Mean	Bias	Mean	Bias
\hat{y}	0	104	-0.5	50	-1.9	54	-1	0	1.2
	0.2	126	-19.6	86	-21	61	0	0	0.9
\hat{y}_v	0	16	-0.1	12	0.3	11	0	-4	-0.2
	0.2	18	-1	16	-1	12	0.1	-4	-0.1

^a The values in the table are actual values divided by $N^2 10^{-4}$

6. DISCUSSION

This paper describes a method for estimating the variance of a survey estimate when some of the values are imputed using hot deck imputation. The method uses a model-assisted approach and conditions on indices for sample members, respondents, and hot deck donors. The approach extends the work of Deville and Särndal (1994) to variance estimation for hot deck imputation, probably the most widely used method of imputation in household surveys. The proposed variance estimator is valid for a general

sample design and for a variety of estimation procedures under the superpopulation model and unconfounded assumptions. The paper also extends the previous work by handling stochastic rather than deterministic imputation and giving conditions for the bias of the naïve variance estimator as an estimator of V'_{SAM} to be small.

The results focus attention on the need to take the mixed component into account when the sample elements have unequal weights. In particular, since domain estimates can be treated by assigning adjusted weights of zero for sample elements not in the domain, the mixed term needs to be taken into account in estimating the variance of domain imputed estimates even if the original weights were equal. Other statistics can also be covered by the approach used for domain estimates. For example, for the simple regression of y on x , with y including hot deck imputed values and x complete, the regression coefficient can be expressed as a weighted linear combination of the y 's: $b = \sum w_i (x_i - \bar{x}) y_i / \sum w_i (x_i - \bar{x})^2 = \sum w'_i y_i$, where $w'_i = w_i (x_i - \bar{x}) / \sum w_i (x_i - \bar{x})^2$. Also the difference between two domain estimates, $\hat{\theta}_{v1}$ and $\hat{\theta}_{v2}$, can be expressed as $\hat{\theta}_{v1} - \hat{\theta}_{v2} = \sum_{i \in v1} w_i y_i - \sum_{i \in v2} w_i y_i = \sum w'_i y_i$, where $w'_i = w_i$ for $i \in v1$, $w'_i = -w_i$ for $i \in v2$, and $w'_i = 0$ for $i \notin v1 \cup v2$.

The last example, involving the difference between domain estimates where imputation cells cut across domains, highlights the importance of the model in the imputation process. In this example, the analytic interest in the difference between the domain statistics is incompatible with an imputation model that assumes no difference in y -distribution across domains within imputation cells. By imputing across domains with a hot deck cell imputation scheme, the sample domain means for y will be brought closer together, thus decreasing the estimate of the difference. Thus, a good imputation model is crucial for producing valid point estimates.

The model-assisted approach to variance estimation with imputed data described here assumes a linear estimator, but smooth nonlinear functions can also be included using a Taylor series approximation. Like the Rao and Shao (1992) adjusted jackknife method, the model-assisted method is applicable with general sample designs and estimation schemes. However, the adjusted jackknife method is applicable only with a weighted hot-deck whereas, as a result of its model assumptions, the model-assisted method can be employed with a variety of hot deck methods, including choosing donors with equal probability and with probabilities proportional to their weights. The model-assisted method of variance estimation could also be extended to other imputation schemes such as nearest neighbor imputation and fractional hot deck imputation (Kalton and Kish 1984; Fay 1996; Kim 2000), a technique which reduces the variance due to imputation.

Implementation of the model-assisted method with hot deck imputation requires the availability of the information needed to compute the three components of the total variance. Standard survey sampling variance estimation software can be used to compute an estimate of \hat{V}_0 that is approximately unbiased with large samples, but as the simulation study illustrates the estimate may be biased if the cell mean model does not hold. The computations of the other components require information on the identity of the donor for each imputed value and of the imputation cell membership of all sample members. From this information, d_{ij} and γ_{ij} can be determined. In addition, an estimate of σ_g^2 is required.

While the theory given above applies to variance estimation with many sample designs, including multi-stage samples, there are serious concerns about the validity of the imputation model in many cases. In the case of multi-stage sampling, the means of many survey variables differ across PSUs, yet hot deck cells are seldom formed within PSUs. Rather they are constructed in terms of other variables that cut across PSUs. Even within these cells there may be differences in means between PSUs. These differences may be offsetting to some extent and not introduce substantial biases for point estimation. However, their effect on variance estimation may be more significant. As indicated in the simulation, failure of the assumptions may have a greater impact on second order statistics than first order statistics. This issue merits more detailed investigation.

Imputation is more difficult when the goal is estimating a function of more than one variable with missing values. To produce an unbiased estimate of a parameter that involves several variables subject to imputation requires the development of an appropriate multivariate model and an imputation procedure consistent with that model. Given an appropriate model and a hot deck imputation that is consistent with it, the model-assisted approach to variance estimation can then be implemented. However, estimating the variance becomes considerably more complex with multivariate estimates. The development of practical methods of imputation and variance estimation for this situation is much needed.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the National Center for Education Statistics, Institute for Education Sciences for funding this research, and in particular the support of the Project Officer, Marilyn Seastrom. We also would like to thank the Associate Editor and referees for their constructive comments.

REFERENCES

- AIGNER, D.J., GOLDBERGER, A.S. and KALTON, G. (1975). On the explanatory power of dummy variable regressions. *International Economic Review*. 16, 503-509.
- BRICK, J.M., and KALTON, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*. 5, 215-238.
- COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 24, 295-313.
- DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87, 376-382.
- DEVILLE, J.C., and SÄRNDAL, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*. 10, 381-394.
- FAY, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*. 91, 490-498.
- FULLER, W.A. (2002). Regression estimation for sample surveys. *Survey Methodology*. 28, 5-23.
- GAGNON, F., LEE, H., RANCOURT, E. and SÄRNDAL, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the Generalized Estimation System. *Proceedings of the Survey Methods Section*, Statistical Society of Canada. 151-156.
- HOLT, D., and SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society. Series A*, 142, Part 1. 33-46.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor: Institute for Social Research, University of Michigan.
- KALTON, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*. 18, 129-154.
- KALTON, G., and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics*. 13(16), 1919-1939.
- KIM, J.K. (2000). Variance estimation after imputation. Unpublished Ph.D. thesis, Iowa State University.
- KIM, J.K., and FULLER, W.A. (1999). Jackknife variance estimation after hot deck imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 825-830.
- LEE, H., RANCOURT, E. and SÄRNDAL, C.-E. (1995). Variance estimation in the presence of imputed data for the Generalized Estimation System. *Proceedings of the Survey Methods Section*, Statistical Society of Canada. 384-389.
- LITTLE, R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*. 54, 139-157.
- RANCOURT, E., SÄRNDAL, C.-E. and LEE, H. (1994). Estimation of the variance in the presence of nearest neighbor imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 888-893.
- RAO, J.N.K. (1999). Some current trends in sample survey theory and methods. *Sankhyā (B)*. 61, 1-57.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 79, 811-822.
- RAO, J.N.K., YUNG, W. and HIDIROGLOU, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā (A)*. 64, 364-378.
- RUBIN, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
- SÄRNDAL, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*. 18, 241-252.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.
- TOLLEFSON, M., and FULLER, W.A. (1992). Variance estimation for samples with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 140-145.

Domain Estimation Using Linear Regression

MICHAEL A. HIDIROGLOU and ZDENEK PATAK¹

ABSTRACT

One of the main objectives of a sample survey is the computation of estimates of means and totals for specific domains of interest. Domains are determined either before the survey is carried out (primary domains) or after it has been carried out (secondary domains). The reliability of the associated estimates depends on the variability of the sample size as well as on the y -variables of interest. This variability cannot be controlled in the absence of auxiliary information for subgroups of the population. However, if auxiliary information is available, the estimated reliability of the resulting estimates can be controlled to some extent. In this paper, we study the potential improvements in terms of the reliability of domain estimates that use auxiliary information. The properties (bias, coverage, efficiency) of various estimators that use auxiliary information are compared using a conditional approach.

KEY WORDS : Domain estimation; Auxiliary data; Conditional properties.

1. INTRODUCTION

One of the main objectives of a sample survey is to compute estimates of means and totals of a number of characteristics associated with the units of a finite population U . The data are often used for analytic studies such as the comparison of means and totals for subgroups of the population. Such subgroups are referred to as *domains of study*. Hartley's (1959) paper is one of the first attempts to unify the theory of domain estimation. Hartley provided the theory for a number of sample designs where domain estimation was of interest. His paper mostly discussed estimators that did not make use of auxiliary information. He did, however, consider the case of the ratio estimator where population totals were known for the domains. The use of auxiliary data in the context of domain estimation has been discussed in a number of articles. Särndal, Swensson and Wretman (1992) provided a unified treatment of domain estimation with auxiliary data. Estevao, Hidiroglou and Särndal (1995) were the first to recognize that the weights accounting for auxiliary data could be domain dependent or not domain dependent. Estevao and Särndal (1999) discussed desirable properties of regression estimators of domain totals using auxiliary data.

The existence of multivariate auxiliary data raises a number of questions in the context of domain estimation. Some of those questions are as follows. What is the effect of having auxiliary information that is not known on a population basis for the given domain of interest? How do we compute valid variance estimates in the context of domain estimators that use auxiliary data? If more than one estimator is possible for point estimation and/or variance estimation, what criteria should be used to choose the best

estimator? Durbin (1969) supported the use of conditional inference to do such comparisons. He stated, "If the sample size is determined by a random mechanism and one happens to get a large sample, one knows perfectly well that the quantities of interest are measured more accurately than they would have been if the sample size had happened to be small. It seems self evident that one should use the information available on sample size in the interpretation of the result. To average over variations in sample size which might have occurred but did not occur, when in fact the sample size is exactly known, seems quite wrong from the standpoint of the analysis of the data actually observed". Holt and Smith (1979) favored conditional inference, and applied it to study the properties of the post-stratified estimator, given simple random sampling. Rao (1985) introduced the idea of "recognizable subsets" of the population to formalize the conditioning process. Recognizable subsets are defined *after* the sample has been drawn. In the case of domain estimation the number of units belonging to a particular domain is a random variable. Recognizable subsets in that context are those where the sample size is fixed within each domain. Comparison of the conditional statistical properties (*i.e.*, bias, mean squared error) of the different estimators can then be based on these subsets. The conditioning process assumes that population totals are known for each domain. In the case of simple random sampling, the number of units in the population domain is assumed known.

The main purpose of this paper is to study the unconditional and conditional properties of a number of domain estimators of totals in the presence of auxiliary data in the context of simple random sampling without replacement (SRSWOR). These conditional properties will be established by conditioning on fixed sample sizes within each domain.

¹ Michael Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; Zdenek Patak, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

The paper is organized as follows. In section 2 we will introduce several estimators of domain totals. Their unconditional and conditional properties are provided in section 3. In section 4, we will present the results of a simulation study for the case of the ratio estimator of domain totals, and provide some concluding remarks in section 5.

2. ESTIMATORS OF DOMAIN TOTALS

We first introduce some notation to set up the framework, under which we will be assessing the performance of various estimators of domain totals. Let $U = \{1, \dots, k, \dots, N\}$ denote the finite population. A sample “ s ” is selected from this population using a sampling plan $P(s)$. Let the first and second order inclusion probabilities be given by π_k and π_{kl} . The domain total $Y_d = \sum_{U_d} y_k$ is the parameter of interest for a variable “ y ”. A domain U_d ($d = 1, \dots, D$) is any subpopulation of U , for which a separate estimate may be required, before or after the planning stage. The number of population units in domain U_d is denoted N_d and $N = \sum_{d=1}^D N_d$ for D mutually exclusive and exhaustive domains spanning the entire population. The sample s is correspondingly divided into D domains $s_1, \dots, s_d, \dots, s_D$ where $s_d = U_d \cap s$. The realized sample size within s_d is a random variable that we denote n_d . Note that the sum of the n_d ’s over non-overlapping and exhaustive domains of the sample equals n . An estimator of the domain total $Y_d = \sum_{U_d} y_k$ that does not use auxiliary data is given by $\hat{Y}_{d,HT} = \sum_{s_d} w_k y_k = \sum_{s_d} w_k y_{dk}$ where $w_k = \pi_k^{-1}$, and y_{dk} is equal to y_k if $k \in U_d$ and 0 otherwise.

Auxiliary information in the form of a p -dimensional vector \mathbf{x} may be available at different levels of aggregation. It may be known for each unit in the population, or for subsets $U_g \subseteq U$ ($g = 1, \dots, G$) of the population U that may coincide with the domains U_d . We denote such known totals $\mathbf{X}_g = \sum_{U_g} \mathbf{x}_k$; they are estimated by $\hat{\mathbf{X}}_{g,HT} = \sum_{s_g} w_k \mathbf{x}_k$. A modified set of weights \tilde{w}_k incorporating the auxiliary data can be computed using either calibration or linear regression procedures (LR). We chose the LR approach. In the case of G population groups, the LR estimator is given by

$$\hat{Y}_{tr} = \hat{Y}_{HT} + \sum_{g=1}^G (\mathbf{X}_g - \hat{\mathbf{X}}_{g,HT})' \hat{\mathbf{B}}_g \quad (1.1)$$

where $\hat{\mathbf{B}}_g = (\sum_{s_g} w_k \mathbf{x}_k \mathbf{x}_k' / c_k)^{-1} \sum_{s_g} w_k \mathbf{x}_k y_k / c_k$, and c_k are suitable positive constants. The use of auxiliary data in the domain context offers a wide range of choices for various levels at which auxiliary totals are used and regression models are constructed. To simplify matters, we assume that $g = 1$ (e.g.: a single group U), yielding the

simple regression estimator $\hat{Y}_{tr} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT})' \hat{\mathbf{B}}$, where $\hat{\mathbf{X}}_{HT} = \sum_s w_k \mathbf{x}_k$.

We consider six estimators for estimating the domain population total Y_d . These estimators are based on whether we use the domain totals \mathbf{X}_d or the population total \mathbf{X} , and whether we construct the regression estimator at the domain or at the population levels. The estimators are categorized into Horvitz-Thompson and “Hájek” types. We provide an example of the *ratio* estimator that is associated with each of these estimators.

2.1 Horvitz-Thompson Type Estimators

Case 1

We assume that the auxiliary information \mathbf{x}_k is available at the population level U , $\mathbf{X} = \sum_U \mathbf{x}_k$ and that the domain specific y_{dk} variables are regressed on \mathbf{x}_k , $k \in U$. The resulting population regression parameter $\mathbf{B}_{1d} = (\sum_U \mathbf{x}_k \mathbf{x}_k' / c_k)^{-1} \sum_U \mathbf{x}_k y_{dk} / c_k$ is estimated by $\hat{\mathbf{B}}_{1d} = (\sum_s w_k \mathbf{x}_k \mathbf{x}_k' / c_k)^{-1} \sum_s w_k \mathbf{x}_k y_{dk} / c_k$ and the resulting estimator of the population total Y_d is

$$\hat{Y}_{d,lr} = \hat{Y}_{d,HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT})' \hat{\mathbf{B}}_{1d}. \quad (2.1)$$

Example: The domain ratio estimator given by $\hat{Y}_{d,RAT} = X \hat{R}_{1d}$, where $\hat{R}_{1d} = \hat{Y}_{d,HT} / \hat{X}_{HT}$. This estimator was first suggested by Hidiroglou (1991), and is discussed in more detail in Estevao *et al.* (1995).

If the auxiliary data totals are available at the domain level, $\mathbf{X}_d = \sum_{U_d} \mathbf{x}_k$, then two possible estimators of Y_d (cases 2 and 3) can be constructed, depending on how the population regression parameter is estimated.

Case 2

The population regression parameter

$$\mathbf{B}_{2d} = \left(\sum_{U_d} \mathbf{x}_k \mathbf{x}_k' / c_k \right)^{-1} \sum_{U_d} (\mathbf{x}_k y_k / c_k)$$

is estimated by regressing y_k on \mathbf{x}_k for each domain U_d separately. Its estimator is given by

$$\hat{\mathbf{B}}_{2d} = \left(\sum_{s_d} w_k \mathbf{x}_k \mathbf{x}_k' / c_k \right)^{-1} \sum_{s_d} (w_k \mathbf{x}_k y_k / c_k),$$

and the resulting regression estimator of a domain total is

$$\hat{Y}_{d,lr_2} = \hat{Y}_{d,HT} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HT})' \hat{\mathbf{B}}_{2d} \quad (2.2)$$

where $\hat{\mathbf{X}}_d = \sum_s w_k \mathbf{x}_{dk}$ with \mathbf{x}_{dk} defined similarly to y_{dk} .

Example: The Horvitz-Thompson post-stratified estimator given by $\hat{Y}_{d,POSTR} = X_d \hat{R}_{2d}$, where $\hat{R}_{2d} = \hat{Y}_{d,HT} / \hat{X}_{d,HT}$.

Case 3

The population regression parameter

$$\mathbf{B}_3 = \left(\sum_U (\mathbf{x}_k \mathbf{x}_k' / c_k) \right)^{-1} \sum_U (\mathbf{x}_k y_k / c_k)$$

is estimated by regressing y_k on \mathbf{x}_k using all units in U . The corresponding estimator is

$$\hat{\mathbf{B}}_3 = \left(\sum_s w_k \mathbf{x}_k \mathbf{x}_k' / c_k \right)^{-1} \sum_s (w_k \mathbf{x}_k y_k / c_k),$$

resulting in the regression estimator

$$\hat{Y}_{d,lr_3} = \hat{Y}_{d,HT} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HT})' \hat{\mathbf{B}}_3. \quad (2.3)$$

Example: The alternate ratio estimator given by $\hat{Y}_{d,ALTR} = \hat{Y}_{d,HT} + (X_d - \hat{X}_{d,HT}) \hat{R}_3$, where $\hat{R}_3 = \hat{Y}_{HT} / \hat{X}_{HT}$.

2.2 Hájek Type Estimators

Estimators (2.1)-(2.3) belong to the Horvitz-Thompson family. If the known population domain size N_d is also incorporated in the estimation, then we get the ‘‘Hájek’’ versions of the previously defined Horvitz-Thompson regression estimators. The Hájek regression estimators are obtained by replacing $\hat{Y}_{d,HT}$, $\hat{\mathbf{X}}_{d,HT}$, and $\hat{\mathbf{X}}_{HT}$ by

$$\hat{Y}_{d,HA} = (N_d / \hat{N}_d) \hat{Y}_{d,HT}, \quad \hat{\mathbf{X}}_{d,HA} = (N_d / \hat{N}_d) \hat{\mathbf{X}}_d,$$

and

$$\hat{\mathbf{X}}_{HA} = (N / \hat{N}) \hat{\mathbf{X}}_{HT},$$

where $\hat{N}_d = \sum_{s_d} w_k$ and $\hat{N} = \sum_s w_k$. The estimators are nearly conditionally unbiased for a given n_d , whereas their Horvitz-Thompson counterparts do not have this property. The ‘‘ $\hat{\mathbf{B}}$ ’’s contained within the Hájek regression estimators correspond exactly to their Horvitz-Thompson counterparts.

Case 4

$$\tilde{Y}_{d,lr_1} = \hat{Y}_{d,HA} + (\mathbf{X} - \hat{\mathbf{X}}_{HA})' \hat{\mathbf{B}}_{1d}. \quad (2.4)$$

Example: The Hájek ratio estimator given by $\tilde{Y}_{d,RAT} = \hat{Y}_{d,HA} + (X - \hat{X}_{HA}) \hat{R}_{1d}$.

Case 5

$$\tilde{Y}_{d,lr_2} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})' \hat{\mathbf{B}}_{2d}. \quad (2.5)$$

Example: The Hájek post-stratified ratio estimator given by $\tilde{Y}_{d,POSTR} = \hat{Y}_{d,HA} + (X_d - \hat{X}_{d,HA}) \hat{R}_{2d}$. This estimator is identical to the Horvitz-Thompson post-stratified estimator.

Case 6

$$\tilde{Y}_{d,lr_3} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})' \hat{\mathbf{B}}_3. \quad (2.6)$$

Example: The Hájek alternate ratio estimator given by $\tilde{Y}_{d,ALTR} = \hat{Y}_{d,HA} + (X_d - \hat{X}_{d,HA}) \hat{R}_3$.

3. PROPERTIES OF THE DOMAIN ESTIMATORS

Estimators (2.1) - (2.6) may be expressed as:

$$\hat{Y}_{d,lr} = \sum_s w_k a_{dk} y_{dk} = \sum_s \tilde{w}_{dk} y_{dk} \quad (2.7)$$

where a_{dk} is an adjustment factor that may or may not be domain dependent. The product of the design weight w_k and the adjustment factor a_{dk} is known as the regression weight (or calibration weight) \tilde{w}_{dk} . Tables 1 and 2 provide a summary of these factors, as well as the residuals required for unconditional variance estimation. The population and sample residuals are denoted as E_{dk} and e_{dk} . The indicator variable δ_{dk} is equal to one if $k \in U_d$ and zero otherwise.

The approximate population and corresponding estimated variances of the Horvitz-Thompson estimators \hat{Y}_{d,lr_j} ($j = 1, 2, 3$) are:

$$V(\hat{Y}_{d,lr_j}) = \sum \sum_U \Delta_{kl} \left(\frac{E_{dk}}{\pi_k} \right) \left(\frac{E_{dl}}{\pi_l} \right) \quad (2.8)$$

and

$$v(\hat{Y}_{d,lr_j}) = \sum \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{a_{dk} e_{dk}}{\pi_k} \right) \left(\frac{a_{dl} e_{dl}}{\pi_l} \right) \quad (2.9)$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$; $\pi_{kl} = \Pr\{k, \ell \in s\}$ with the appropriate E_{dk} 's, e_{dk} 's, a_{dk} 's defined in Table 1.

The approximate unconditional population and corresponding estimated variances of the Hájek-type estimators \tilde{Y}_{d,lr_j} ($j = 1, 2, 3$) are:

$$V(\tilde{Y}_{d,lr_j}) = \begin{cases} \sum \sum_U \Delta_{kl} \left(\frac{E_{dk} - \left(\sum_U E_{dk} / N_d \right) \delta_{dk}}{\pi_k} \right) \times \\ \left(\frac{E_{dl} - \left(\sum_U E_{dk} / N_d \right) \delta_{dl}}{\pi_l} \right) & \text{for } j = 1 \\ \sum \sum_{U_d} \Delta_{kl} \left(\frac{E_{dk} - \tilde{E}_{U_d}}{\pi_k} \right) \left(\frac{E_{dl} - \tilde{E}_{U_d}}{\pi_l} \right) & \text{for } j = 2, 3 \end{cases} \quad (2.10)$$

Table 1
Adjustment Factors and Residuals for Horvitz-Thompson Regression Estimators

Estimator	Domain Dependent	Adjustment Factor: a_{dk}	Residuals
$\hat{Y}_{d,\ell r_1}$	No	$1 + (\mathbf{X} - \hat{\mathbf{X}}_{\text{HT}})' \left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$
$\hat{Y}_{d,\ell r_2}$	Yes	$\delta_{dk} \left(1 + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\text{HT}})' \left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k} \right)$	$E_{dk} = y_{dk} - \mathbf{x}_k' \mathbf{B}_{2d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{2d}$
$\hat{Y}_{d,\ell r_3}$	Yes	$\delta_{dk} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\text{HT}})' \left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \mathbf{B}_3$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_3$

Table 2
Adjustment Factors and Residuals for the Hájek-type Estimators

Estimator	Domain Dependent	Adjustment Factor: a_{dk}	Residuals
$\tilde{Y}_{d,\ell r_1}$	Yes	$\frac{N_d}{\hat{N}_d} + (\mathbf{X} - \hat{\mathbf{X}}_{\text{HA}})' \left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \mathbf{B}_{1d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$
$\tilde{Y}_{d,\ell r_2}$	Yes	$\delta_{dk} \left(\frac{N_d}{\hat{N}_d} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\text{HA}})' \left(\sum_{s_d} \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k} \right)$	$E_{dk} = y_{dk} - \mathbf{x}_k' \mathbf{B}_{2d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{2d}$
$\tilde{Y}_{d,\ell r_3}$	Yes	$\delta_{dk} \frac{N_d}{\hat{N}_d} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,\text{HA}})' \left(\sum_s \frac{w_k \mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \mathbf{B}_3$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_3$

and

$$v(\tilde{Y}_{d,\ell r_j}) = \sum \sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}} \left(\frac{a_{dk} e_{dk}}{\pi_k} \right) \left(\frac{a_{d\ell} e_{d\ell}}{\pi_\ell} \right) \quad \text{for } j = 1, 2, 3 \quad (2.11)$$

where $\tilde{E}_{U_d} = \sum_{U_d} E_{dk} / N_d$. The appropriate E_{dk} 's, e_{dk} 's, and a_{dk} 's are defined in Table 2. Note that the form of the estimated unconditional variance is the same for both the Horvitz-Thompson and the Hájek-type estimators.

Result 3.1: The Hájek-type regression estimator can be obtained as a by-product of the regression of y_k on

$$(\mathbf{x}_k^\circ)' = \left(1, (\mathbf{x}_k - \bar{\mathbf{x}}_U)' \right),$$

where $\bar{\mathbf{x}}_U = N^{-1} \sum_U \mathbf{x}_k$. The resulting regression vector is

$$\hat{\mathbf{B}}^\circ = \left(\hat{B}_1^\circ, \hat{\mathbf{B}}_x^\circ \right),$$

where

$$\hat{\mathbf{B}}_x^\circ = \left(\left(\sum_s w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s) (\mathbf{x}_k - \tilde{\mathbf{x}}_s)' / c_k \right)^{-1} \times \sum_s (w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s) y_k / c_k) \right)$$

and $\hat{B}_1^\circ = \tilde{y}_s + (\bar{\mathbf{x}}_U - \tilde{\mathbf{x}}_s)' \hat{\mathbf{B}}_x^\circ$, with $\tilde{y}_s = \hat{Y}_{\text{HT}} / \hat{N}$ and $\tilde{\mathbf{x}}_s = \hat{\mathbf{X}}_{\text{HT}} / \hat{N}$.

The regression estimator of total $\hat{Y}_{\ell r} = N \hat{B}_1^\circ$ is equal to the Hájek form $\hat{Y}_{\ell r} = \hat{Y}_{\text{HA}} + (\mathbf{X} - \hat{\mathbf{X}}_{\text{HA}})' \hat{\mathbf{B}}_x^\circ$. The various Hájek-type domain regression estimators can be obtained using this approach. For instance, regressing y_{dk} on

$$(\mathbf{x}_k^\circ)' = \left(1, (\mathbf{x}_{dk} - \bar{\mathbf{x}}_{U_d})' \right)$$

yields $\tilde{Y}_{d,\ell r_1}$.

Proof: We first show how to arrive at the Hájek form of the regression estimator. Defining the auxiliary data vector \mathbf{z}_k as $\mathbf{z}_k' = (x_{0k}, \mathbf{x}_k)'$, the regression estimator is

$$\hat{Y}_{\ell r} = \hat{Y}_{\text{HT}} + (\mathbf{Z} - \hat{\mathbf{Z}}_{\text{HT}})' \hat{\mathbf{B}}_z$$

where

$$\hat{\mathbf{B}}_z = \left(\sum_s w_k \mathbf{z}_k \mathbf{z}_k' / c_k \right)^{-1} \left(\sum_s w_k \mathbf{z}_k y_k / c_k \right),$$

$\mathbf{Z} = \sum_U \mathbf{z}_k$ and $\hat{\mathbf{Z}}_{\text{HT}} = \sum_U w_k \mathbf{z}_k$.

If $x_{0k} = 1$, $\hat{Y}_{\ell r}$ is exactly equivalent to $\hat{Y}_{\ell r} = \mathbf{Z}' \hat{\mathbf{B}}_z$. Decomposing $\hat{\mathbf{B}}_z$ as

$$\hat{\mathbf{B}}_z' = \left(\hat{B}_0, \hat{\mathbf{B}}_x' \right),$$

we have that $\hat{Y}_{\ell r} = N \hat{B}_0 + \sum_U \mathbf{x}_k' \hat{\mathbf{B}}_x$, where $\hat{B}_0 = \tilde{y}_s - \tilde{\mathbf{x}}_s' \hat{\mathbf{B}}_x$ and

$$\hat{\mathbf{B}}_x = \left(\frac{\sum_s w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s) (\mathbf{x}_k - \tilde{\mathbf{x}}_s)'}{c_k} \right)^{-1} \times \frac{\sum_s w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s) (y_k - \tilde{y}_s)}{c_k}.$$

Hence, the Hájek form of the regression estimator is

$$\tilde{Y}_{\ell r} = \hat{Y}_{HA} + (\mathbf{X}_U - \hat{\mathbf{X}}_{HA})' \hat{\mathbf{B}}_x.$$

Regressing y_k on

$$(\mathbf{x}_k^*)' = \left(1, (\mathbf{x}_k - \bar{\mathbf{x}}_U)'\right)$$

yields the estimated regression vector $\hat{\mathbf{B}}_x' = (\hat{B}_1^*, \hat{B}_x^*)'$, where

$$\hat{\mathbf{B}}_x = \left(\frac{\sum_s w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s)(\mathbf{x}_k - \tilde{\mathbf{x}}_s)'}{c_k} \right)^{-1} \sum_s \frac{w_k (\mathbf{x}_k - \tilde{\mathbf{x}}_s) y_k}{c_k}$$

and $\hat{B}_1^* = \bar{y}_s + (\bar{\mathbf{x}}_U - \tilde{\mathbf{x}}_s)' \hat{\mathbf{B}}_x$. Substituting \hat{B}_1^* into $\hat{Y}_{\ell r} = N \hat{B}_1^0$ yields the Hájek form $\tilde{Y}_{\ell r}$.

Remark 3.1: (Additivity). Suppose that the domains U_d are mutually exclusive ($U_{d_1} \cap U_{d_2} = \emptyset$ for $d_1 \neq d_2$) and exhaustive ($\bigcup_{d=1}^D U_d = U$). Additivity over such domains means that $\sum_{d=1}^D \hat{Y}_{d, \ell r_1} = \sum_{d=1}^D \hat{Y}_{d, \ell r_2} = \hat{Y}_{\ell r}$ where

$$\hat{Y}_{\ell r} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT})' \hat{\mathbf{B}}.$$

The additive property of $\hat{Y}_{d, \ell r_1}$ is desirable because a single set of calibration weights, w_k , can be used repeatedly to produce ad hoc domain estimates. Only two out of the six estimators, $\hat{Y}_{d, \ell r_1}$ and $\hat{Y}_{d, \ell r_3}$, are additive over all such domains.

Remark 3.2: (Calibrating on domain auxiliary data). Esteveo *et al.* (1999) discussed some of the estimators provided in Tables 1 and 2 for the case of a single auxiliary variable x_k . They arrived at their estimators by controlling on domain information, either via auxiliary variables and/or control totals.

In what follows, we will assume that the sample s of size n has been selected using simple random sampling without replacement (SRSWOR) from a universe of size N . The estimated unconditional variance of the Horvitz-Thompson and Hájek-type estimators for this sampling plan is:

$$v(\hat{Y}_{d, \ell r_1}) = \sum_s \frac{N^2(1-f)}{n} \frac{\sum_s (a_{dk} e_{dk} - \overline{a_d e})^2}{n-1} \quad (2.12)$$

where $\overline{a_d e} = \sum_s (a_{dk} e_{dk} / n)$ and $f = n/N$ is the sampling fraction.

3.1 Unconditional Properties

The choice between the various regression estimators should be based on the level at which the auxiliary totals are available, as well as bias and variance. All the above estimators are asymptotically unconditionally unbiased; however, their variances differ. We compare the unconditional

population variances of the six domain regression estimators (2.1) – (2.6) by distinguishing two cases: (i) an intercept term is included in the regression; and (ii) no intercept term is included in the regression.

Result 3.2: Assume that an intercept is included in the regression, $c_k = c$ for all $k \in U$, and $N > p$, where p refers to the number of auxiliary variables. The following inequalities hold for the population variances of the domain regression estimators (2.1) – (2.6):

- (i) $V(\hat{Y}_{d, \ell r_2}) < V(\hat{Y}_{d, \ell r_1})$; $V(\hat{Y}_{d, \ell r_2}) \leq V(\hat{Y}_{d, \ell r_3})$; $V(\hat{Y}_{d, \ell r_3})$ may be smaller, equal or greater to $V(\hat{Y}_{d, \ell r_1})$.
- (ii) $V(\tilde{Y}_{d, \ell r_2}) < V(\tilde{Y}_{d, \ell r_1})$ and $V(\tilde{Y}_{d, \ell r_2}) < V(\tilde{Y}_{d, \ell r_3})$; $V(\tilde{Y}_{d, \ell r_3})$ may be smaller, equal or greater to $V(\tilde{Y}_{d, \ell r_1})$.

Proof. In the case of simple random sampling without replacement, $V(\hat{Y}_{d, \ell r_1}) = A \sum_U (E_{dk} - \bar{E}_{U_d})^2$ for $\ell = 1, 2, 3$, where $A = N^2(1-f)/n(N-1)$ and $\bar{E}_{U_d} = \sum_{U_d} E_{dk} / N$. Given that the regression contains an intercept, it follows that $\sum_U E_{dk} = 0$ or that $\sum_{U_d} E_{dk} = 0$, depending on which regression estimator we use. We only show that (i) holds: the proof for (ii) is similar. The population variances for $\hat{Y}_{d, \ell r_1}$ and $\hat{Y}_{d, \ell r_2}$ are respectively

$$V(\hat{Y}_{d, \ell r_1}) = A \sum_U (y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d})^2$$

and

$$V(\hat{Y}_{d, \ell r_2}) = A \sum_U (y_{dk} - \mathbf{x}'_k \mathbf{B}_{2d})^2.$$

The population variance of $\hat{Y}_{d, \ell r_3}$ is

$$V(\hat{Y}_{d, \ell r_3}) = A \sum_U (E_{dk} - \bar{E}_{U_d})^2,$$

where

$$\bar{E}_{U_d} = N^{-1} \sum_U E_{dk} = \left(\frac{N_d}{N} \right) (\bar{y}_{U_d} - \bar{\mathbf{x}}'_{U_d} \mathbf{B}_3),$$

with $\bar{y}_{U_d} = N_d^{-1} \sum_{U_d} y_{dk}$ and $\bar{\mathbf{x}}'_{U_d}$ similarly defined.

We first show that $V(\hat{Y}_{d, \ell r_2}) < V(\hat{Y}_{d, \ell r_1})$. To this end, we decompose $\sum_U (y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d})^2$ into its within domain U_d and outside domain U_d components, yielding

$$\begin{aligned} \sum_U (y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d})^2 &= \sum_{U_d} (y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d})^2 \\ &\quad + \sum_{U_d^c} (y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d})^2. \end{aligned}$$

Since

$$\begin{aligned} \sum_{U_d} (y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d})^2 &= \sum_{U_d} (y_{dk} - \mathbf{x}'_k \mathbf{B}_{2d})^2 \\ &\quad + \sum_{U_d} (\mathbf{x}'_k (\mathbf{B}_{2d} - \mathbf{B}_{1d}))^2, \end{aligned}$$

it follows that $V(\hat{Y}_{d, \ell r_2}) < V(\hat{Y}_{d, \ell r_1})$.

Next, we show that

$$V(\hat{Y}_{d, \ell r_2}) \leq V(\hat{Y}_{d, \ell r_3}).$$

The variance $V(\hat{Y}_{d, \ell r_3})$ can be re-expressed as

$$V(\hat{Y}_{d,\ell r_3}) = \sum_{U_d} \begin{pmatrix} (y_k - \mathbf{x}'_k \mathbf{B}_3)^2 \\ - \left(\frac{N_d^2}{N} \right) (\bar{y}_{U_d} - \bar{\mathbf{x}}'_{U_d} \mathbf{B}_3)^2 \end{pmatrix}.$$

The difference between $V(\hat{Y}_{d,\ell r_3})$ and $V(\hat{Y}_{d,\ell r_2})$ is:

$$\begin{aligned} & V(\hat{Y}_{d,\ell r_3}) - V(\hat{Y}_{d,\ell r_2}) \\ &= A \left\{ \sum_{U_d} \left((y_k - \mathbf{x}'_k \mathbf{B}_3)^2 \right) - \left(\frac{N_d^2}{N} \right) (\bar{y}_{U_d} - \bar{\mathbf{x}}'_{U_d} \mathbf{B}_3)^2 \right. \\ &\quad \left. - \sum_{U_d} (y_k - \mathbf{x}'_k \mathbf{B}_{2d})^2 \right\} \\ &= A \left\{ (\mathbf{B}_3 - \mathbf{B}_{2d})' \left(\sum_{U_d} \mathbf{x}_k \mathbf{x}'_k \right) (\mathbf{B}_3 - \mathbf{B}_{2d}) \right. \\ &\quad \left. - \left(\frac{N_d^2}{N} \right) (\bar{y}_{U_d}^2 - 2\mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{y}_{U_d} + \mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} \mathbf{B}_3) \right\} \\ &\geq A \left\{ (\mathbf{B}_3 - \mathbf{B}_{2d})' \left(\sum_{U_d} \mathbf{x}_k \mathbf{x}'_k \right) (\mathbf{B}_3 - \mathbf{B}_{2d}) \right. \\ &\quad \left. - N_d (\bar{y}_{U_d}^2 - 2\mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{y}_{U_d} + \mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} \mathbf{B}_3) \right\}. \end{aligned}$$

Noting that $\bar{y}_{U_d} = \bar{\mathbf{x}}'_{U_d} \mathbf{B}_{2d}$ it follows that:

$$\begin{aligned} & \bar{y}_{U_d}^2 - 2\mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{y}_{U_d} + \mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} \mathbf{B}_3 \\ &= \mathbf{B}_{2d}' \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} \mathbf{B}_{2d} - 2\mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} \mathbf{B}_{2d} \\ &\quad + \mathbf{B}_3' \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} \mathbf{B}_3 \\ &= (\mathbf{B}_3 - \mathbf{B}_{2d})' \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} (\mathbf{B}_3 - \mathbf{B}_{2d}) \end{aligned}$$

Since

$$\sum_{U_d} \mathbf{x}_k \mathbf{x}'_k - N_d \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} = \sum_{U_d} (\mathbf{x}_k - \bar{\mathbf{x}}_{U_d}) (\mathbf{x}_k - \bar{\mathbf{x}}_{U_d})',$$

the difference $V(\hat{Y}_{d,\ell r_3}) - V(\hat{Y}_{d,\ell r_2})$ can be expressed as:

$$\begin{aligned} & V(\hat{Y}_{d,\ell r_3}) - V(\hat{Y}_{d,\ell r_2}) \\ &= A \left\{ (\mathbf{B}_3 - \mathbf{B}_{2d})' \left(\sum_{U_d} \mathbf{x}_k \mathbf{x}'_k - N_d \bar{\mathbf{x}}_{U_d} \bar{\mathbf{x}}'_{U_d} \right) (\mathbf{B}_3 - \mathbf{B}_{2d}) \right\} \\ &= A \left\{ (\mathbf{B}_3 - \mathbf{B}_{2d})' \sum_{U_d} (\mathbf{x}_k - \bar{\mathbf{x}}_{U_d}) (\mathbf{x}_k - \bar{\mathbf{x}}_{U_d})' (\mathbf{B}_3 - \mathbf{B}_{2d}) \right\} \\ &\geq 0. \end{aligned}$$

Finally, we show that $V(\hat{Y}_{d,\ell r_3})$ may be smaller, equal or greater to $V(\hat{Y}_{d,\ell r_1})$ by constructing examples:

- (i) $V(\hat{Y}_{d,\ell r_3}) < V(\hat{Y}_{d,\ell r_1})$, if $\mathbf{B}_3 = \mathbf{B}_{2d}$;
- (ii) $V(\hat{Y}_{d,\ell r_3}) = V(\hat{Y}_{d,\ell r_1})$, if $\mathbf{B}_3 = \mathbf{B}_{1d}$;
- (iii) $V(\hat{Y}_{d,\ell r_3}) > V(\hat{Y}_{d,\ell r_1})$, if the fit of y_k on \mathbf{x}_k is much poorer than the fit y_{dk} on \mathbf{x}_k for $k \in U$.

It can also be shown that $V(\tilde{Y}_{d,\ell r_3}) < V(\hat{Y}_{d,\ell r_3})$; $V(\tilde{Y}_{d,\ell r_3}) < V(\hat{Y}_{d,\ell r_1})$; and $V(\tilde{Y}_{d,\ell r_1}) < V(\hat{Y}_{d,\ell r_1})$. The estimator with the smallest variance is $\tilde{Y}_{d,\ell r_3}$. However, if it is assumed that the $\hat{\mathbf{B}}_{2d}$'s are similar across all domains, and that there are very few observations in s_d , it may be preferable to use $\tilde{Y}_{d,\ell r_3}$. The choice between $\tilde{Y}_{d,\ell r_3}$ and

$\tilde{Y}_{d,\ell r_3}$ should not always be based on the asymptotic variance. If there are very few observations in s_d , this can cause significant bias in $\tilde{Y}_{d,\ell r_3}$ and also cause the exact variance of $\tilde{Y}_{d,\ell r_3}$ to be larger than that of $\tilde{Y}_{d,\ell r_1}$, so that the latter may be preferred.

Remark 3.3: If there is no intercept in the regression, then it does not necessarily follow that Result 3.2 holds.

Proof: We illustrate this statement using the elementary ratio versions of cases 1 and 2. They are respectively the *Horvitz-Thompson ratio* estimator $\hat{Y}_{d,\text{RAT}} = \hat{Y}_{d,\text{HT}} (X / \hat{X}_{\text{HT}})$ and the *Horvitz-Thompson post-stratified ratio* estimator $\hat{Y}_{d,\text{POSTR}} = \hat{Y}_{d,\text{HT}} (X_d / \hat{X}_{d,\text{HT}})$. Also, suppose that the elements of the data vector (y_k, x_k) are positive for all $k \in U$. The population variances for $\hat{Y}_{d,\text{RAT}}$ and $\hat{Y}_{d,\text{POSTR}}$ are $V(\hat{Y}_{d,\text{POSTR}}) = A \sum_{U_d} (y_k - B_{2d} x_k)^2$ and $V(\hat{Y}_{d,\text{RAT}}) = A \sum_U (y_{dk} - B_{1d} x_k)^2$, where $B_{2d} = Y_d / X_d$, and $B_{1d} = Y_d / X$.

The difference $V(\hat{Y}_{d,\text{RAT}}) - V(\hat{Y}_{d,\text{POSTR}})$ can be re-expressed as:

$$\begin{aligned} & A \sum_{U_d} (B_{1d} - B_{2d})^2 x_k^2 \\ & + 2A(B_{1d} - B_{2d}) \sum_{U_d} (y_k - B_{2d} x_k) x_k \\ & + A \sum_{U_d} (y_{dk} - B_{1d} x_k)^2. \end{aligned}$$

Since the second term of this expression can be positive, negative or zero, the difference $V(\hat{Y}_{d,\text{RAT}}) - V(\hat{Y}_{d,\text{POSTR}})$ can be negative.

3.2 Conditional Properties

For a given sample s , let n_d be the realized sample size of s_d . The following result can be used to evaluate the conditional bias of estimators (2.1) to (2.6).

Result 3.3: Let \mathbf{z}_k be an arbitrary p -dimensional vector, that is $\mathbf{z}_k = (z_{k1}, \dots, z_{kp})'$, and suppose that $n_d \geq 1$. The conditional expectation of $\bar{\mathbf{z}}_s = n^{-1} \sum_s \mathbf{z}_k$ given n_d can be written as:

$$E(\bar{\mathbf{z}}_s | n_d) = \frac{1}{n} \left[f_d \sum_{U_d} \mathbf{z}_k + f_{\bar{d}} \left(\sum_U \mathbf{z}_k - \sum_{U_d} \mathbf{z}_k \right) \right]$$

$$= \bar{\mathbf{z}}_U + \frac{w_d - W_d}{1 - W_d} (\bar{\mathbf{z}}_{U_d} - \bar{\mathbf{z}}_U) \quad (3.1)$$

where $\bar{\mathbf{z}}_U = N^{-1} \sum_U \mathbf{z}_k$, $\bar{\mathbf{z}}_{U_d} = N_d^{-1} \sum_{U_d} \mathbf{z}_k$, $w_d = n_d / n$, $W_d = N_d / N$, $f_d = n_d / N_d$, $f_{\bar{d}} = n_{\bar{d}} / N_{\bar{d}}$ with $n_{\bar{d}} = n - n_d$, and $N_{\bar{d}} = N - N_d$.

Proof: Rewriting $\bar{\mathbf{z}}_s$ as

$$\frac{1}{n} \left(\sum_{s_d} \mathbf{z}_k + \sum_{s_{\bar{d}}} \mathbf{z}_k \right),$$

we have that

$$E(\bar{\mathbf{z}}_s | n_d) = \frac{1}{n} \left[\frac{n_d}{N_d} \sum_{U_d} \mathbf{z}_k + \frac{n - n_d}{N - N_d} \sum_{U_{\bar{d}}} \mathbf{z}_k \right]$$

where $s_d = \{k \in s \text{ and } k \notin s_d\}$ and

$$U_d = \{k \in U \text{ and } k \notin U_d\}.$$

Since $\sum_{U_d} \mathbf{z}_k = \sum_U \mathbf{z}_k - \sum_{U_d} \mathbf{z}_k$, we obtain the required result, that is

$$E(\bar{\mathbf{z}}_s | n_d) = \bar{\mathbf{z}}_U + \frac{w_d - W_d}{1 - W_d} (\bar{\mathbf{z}}_{U_d} - \bar{\mathbf{z}}_U).$$

Result 3.4: The conditional population variance of $\bar{\mathbf{z}}_s$ given n_d , can be written as

$$V(\bar{\mathbf{z}}_s | n_d) = \frac{w_d^2}{n_d} (1 - f_d) \mathbf{V}_{z_{U_d}} + \frac{w_d^2}{n_d} (1 - f_d) \mathbf{V}_{z_{U_d}},$$

where

$$\mathbf{V}_{z_{U_d}} = \frac{1}{N_d - 1} \sum_{U_d} (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d}) (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d})',$$

$$\mathbf{V}_{z_{U_d}} = \frac{1}{N_d - 1} \sum_{U_d} (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d}) (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d})',$$

with $\bar{\mathbf{z}}_{U_d} = N_d^{-1} \sum_{U_d} \mathbf{z}_k$, and $w_d = 1 - w_d$.

The estimator of the conditional population variance $V(\bar{\mathbf{z}}_s | n_d)$ is given by

$$\mathbf{v}(\bar{\mathbf{z}}_s | n_d) = \frac{w_d^2}{n_d} f_d \mathbf{v}_{z_{U_d}} + \frac{w_d^2}{n_d} (1 - f_d) \mathbf{v}_{z_{U_d}},$$

where

$$\mathbf{v}_{z_{U_d}} = \frac{1}{n_d - 1} \sum_{U_d} (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d}) (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d})'$$

and

$$\mathbf{v}_{z_{U_d}} = \frac{1}{n_d - 1} \sum_{U_d} (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d}) (\mathbf{z}_k - \bar{\mathbf{z}}_{U_d})',$$

with $\bar{\mathbf{z}}_{U_d} = n_d^{-1} \sum_{U_d} \mathbf{z}_k$, $\bar{\mathbf{z}}_{U_d} = n_d^{-1} \sum_{U_d} \mathbf{z}_k$.

Proof. It follows using arguments similar to those used in Result 3.3. We first illustrate how Result 3.3 can be used to obtain the conditional bias for the simpler estimators of domain totals. This includes the Horvitz-Thompson estimator $\hat{Y}_{d,HT}$, as well as post-stratified ratio estimator $\hat{Y}_{d,POSTR} = (X_d / \hat{X}_{d,HT}) \hat{Y}_{d,HT}$. Let \mathbf{z}_k be the domain variable y_{dk} . Using Result 3.3, we have that $E(\hat{Y}_{d,HT} | n_d) = N w_d \bar{y}_{U_d}$, where $\bar{y}_{U_d} = Y_d / N_d$. The conditional bias of $\hat{Y}_{d,HT}$ given n_d is therefore $\text{Bias}(\hat{Y}_{d,HT} | n_d) = N(w_d - W_d) \bar{y}_{U_d}$. For the post-stratified ratio estimator, note that $\hat{Y}_{d,POSTR} - Y_d = \hat{Y}_{d,HT} - (Y_d / X_d) \hat{X}_{d,HT}$. Defining \mathbf{z}_k as $y_{dk} - (Y_d / X_d) x_{dk}$, we obtain that $\text{Bias}(\hat{Y}_{d,POSTR} | n_d) = 0$.

We next proceed to evaluate the conditional bias and variance of estimators (2.1) – (2.6). We only illustrate the procedure for the regression estimator $\hat{Y}_{d,\ell\eta}$, as the steps are similar for the other estimators. Conditional on n_d , the distribution of s_d is that of an SRSWOR. This means that, for each sample s_d , n_d can be considered as having been selected from N_d . We express $\hat{Y}_{d,\ell\eta}$ as $\hat{Y}_{d,\ell\eta} = \sum_U \hat{y}_k + N/n \sum_s e_{dk}$, where $e_{dk} = y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d}$ and $\hat{y}_k = \mathbf{x}'_k \hat{\mathbf{B}}_{1d}$.

Following Särndal and Hidiroglou (1989), we define the conditional regression vector \mathbf{B}_{1d}^* as

$$\mathbf{B}_{1d}^* = \left[E \left(\sum_s \frac{\mathbf{x}_k \mathbf{x}'_k}{c_k} \middle| n_d \right)^{-1} \times E \left[\left(\sum_s \frac{\mathbf{x}_k y_k}{c_k} \right) \middle| n_d \right] \right]. \quad (3.2)$$

The estimated regression vector $\hat{\mathbf{B}}_{1d}$ will converge to \mathbf{B}_{1d}^* (under appropriate conditions) in conditional design probability as n_d and N_d increase.

We have that

$$E \left[\sum_s \frac{\mathbf{x}_k \mathbf{x}'_k}{n c_k} \middle| n_d \right] = \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{N c_k} + \mathbf{R}_c$$

and

$$E \left[\sum_s \frac{\mathbf{x}_k y_k}{n c_k} \middle| n_d \right] = \sum_U \frac{\mathbf{x}_k y_k}{N c_k} + \mathbf{r}_c,$$

where

$$\mathbf{R}_c = \frac{w_d - W_d}{1 - W_d} \left(\frac{1}{N_d} \sum_{U_d} \frac{\mathbf{x}_k \mathbf{x}'_k}{c_k} - \frac{1}{N} \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{c_k} \right) \doteq \mathbf{0}$$

and

$$\mathbf{r}_c = \frac{w_d - W_d}{1 - W_d} \left(\frac{1}{N_d} \sum_{U_d} \frac{\mathbf{x}_k y_k}{c_k} - \frac{1}{N} \sum_U \frac{\mathbf{x}_k y_k}{c_k} \right) \doteq \mathbf{0}.$$

Consequently, using Result 3.3 and assuming that $(w_d - W_d)/(1 - W_d) \doteq 0$, we have that $\hat{\mathbf{B}}_{1d} \doteq \mathbf{B}_{1d}^*$.

Define the “conditional residual” for the k^{th} unit as

$$E_{dk}^* = y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d}^*. \quad (3.3)$$

The deviation of $\hat{Y}_{d,\ell\eta}$ from the true value Y_d can be written as

$$\hat{Y}_{d,\ell\eta} - Y_d = - \sum_U E_{dk}^* + \frac{N}{n} \sum_s E_{dk}^* - \Delta_{1d}^* \quad (3.4)$$

where

$$\Delta_{1d}^* = \left(\frac{N}{n} \sum_s \mathbf{x}_k - \sum_U \mathbf{x}_k \right)' (\hat{\mathbf{B}}_{1d} - \mathbf{B}_{1d}^*).$$

In equation (3.4), Δ_{1d}^* is of lower order than $N/n \sum_s E_{dk}^*$. To see this, note that

$$E \left[\left(\frac{N}{n} \sum_s \mathbf{x}_k - \sum_U \mathbf{x}_k \right) \middle| n_d \right] = N \frac{w_d - W_d}{1 - W_d} (\bar{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U),$$

where $(w_d - W_d)/(1 - W_d)$ should be close to zero.

Also, as noted earlier, $\hat{\mathbf{B}}_{1d} - \mathbf{B}_{1d}^*$ is near the vector $\mathbf{0}$ in conditional design probability. Hence $E_{dk}^* = y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d}^* \doteq y_{dk} - \mathbf{x}'_k \mathbf{B}_{1d} = E_{dk}$. This implies that we can write (3.4) as

$$\hat{Y}_{d,\ell\eta} - Y_d \doteq - \sum_U E_{dk} + \frac{N}{n} \sum_s E_{dk}. \quad (3.5)$$

The conditional expectation of $\hat{Y}_{d,\ell\eta} - Y_d$ is approximately:

$$E[(\hat{Y}_{d,\ell_1} - Y_d) | n_d] = N \frac{w_d - W_d}{1 - W_d} (\tilde{E}_{U_d} - \bar{E}_{U_d}) \quad (3.6)$$

where $\tilde{E}_{U_d} = \sum_{U_d} E_{dk} / N_d$ and $\bar{E}_{U_d} = \sum_U E_{dk} / N$. Since $\bar{Y}_{U_d} = W_d \tilde{Y}_{U_d}$, the conditional expectation (3.6) can be re-expressed as:

$$\begin{aligned} E[(\hat{Y}_{d,\ell_1} - Y_d) | n_d] &= N \frac{w_d - W_d}{1 - W_d} \\ &\quad \left[\tilde{Y}_{U_d} (1 - W_d) - (\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U)' \mathbf{B}_{1d} \right]. \quad (3.7) \end{aligned}$$

The term $\sum_U E_{dk}$ is constant in (3.5). Using Result 3.4, the conditional population variance of \hat{Y}_{d,ℓ_1} and its estimated value are respectively

$$V(\hat{Y}_{d,\ell_1} | n_d) = N^2 \left[\frac{w_d^2}{n_d} (1 - f_d) V_{E_{U_d}} + \frac{w_d^2}{n_d} (1 - f_d) V_{E_{U_d}} \right]$$

and

$$v(\hat{Y}_{d,\ell_1} | n_d) = N^2 \left[\frac{w_d^2}{n_d} (1 - f_d) v_{e_{sd}} + \frac{w_d^2}{n_d} (1 - f_d) v_{e_{\bar{s}}}$$

where $V_{E_{U_d}} = (N_d - 1)^{-1} \sum_{U_d} (E_{dk} - \tilde{E}_{U_d})^2$,

$$V_{E_{U_d}} = (N_d - 1)^{-1} \sum_{U_d} (E_{dk} - \tilde{E}_{U_d})^2,$$

$$v_{e_{sd}} = (n_d - 1)^{-1} \sum_{s_d} \left(a_{dk} e_k - \sum_{s_d} \frac{a_{dk} e_k}{n_d} \right)^2,$$

and

$$v_{e_{\bar{s}}} = (n_d - 1)^{-1} \sum_{s_d} \left(a_{dk} e_{dk} - \sum_{s_d} \frac{a_{dk} e_{dk}}{n_d} \right)^2.$$

The conditional bias and variances of the remaining five estimators can be derived similarly. Table 3 presents a summary of these properties. The required adjustment factors a_{dk} and residual terms e_{dk} are given in Tables 1 and 2.

4. SIMULATION STUDY

A simulation study was carried out to illustrate the conditional and unconditional properties of the ratio version of estimators (2.1) – (2.6). We studied these properties using a population of 1,000 bivariate observations (y, x) . This population resulted from the concatenation of two generated population domains: a large domain of size 900 and a small domain of size 100. The (y, x) observations were generated within each domain assuming a ratio model $y_k = \beta x_k + \varepsilon_k$ where $E(\varepsilon_k) = 0$ and $V(\varepsilon_k) = \sigma^2 x_k$. The β coefficients were 1.0 and 3.0 in the large and small domains. The auxiliary variable x was generated using a gamma distribution $\Gamma(a, b)$, where $a = 3$ and $b = 16$. The dependent variable y was also generated by a gamma distribution, $\Gamma(A, B)$ such that the parameters A and B satisfied $E(y_k) = \beta x_k = AB$ and $V(y_k) = \sigma^2 x_k = AB^2$. After solving for A and B , we obtained $A = \beta^2 / \sigma^2$ and $B = \sigma^2 / \beta$. The term σ^2 was chosen to satisfy a set correlation between x and y defined by

$$\rho_{x,y} = \frac{\beta b}{\sqrt{\sigma^2 b + \beta^2 b^2}}.$$

The preceding equation yields the constant term

$$\sigma^2 = \beta^2 b \left(\frac{1}{\rho_{x,y}^2} - 1 \right)$$

of the error variance. Common correlation values $\rho_{x,y}$ were used for both domains, ranging from 0.1 to 0.9 in steps of 0.1, resulting in nine different populations. Random samples ($M = 10,000$) of size 250 were then repeatedly selected from the populations. For each sample, estimates of domain totals were computed using the estimators given in Table 4. We do not include the Hájek post-stratified estimator, \hat{Y}_{d,ℓ_2} , as it corresponds exactly to its Horvitz-Thompson analogue, \hat{Y}_{d,ℓ_2} .

Table 3
Conditional Bias and Variance of Estimators (2.1)–(2.6)

Estimator	Conditional Bias	Estimated Conditional Variance
\hat{Y}_{d,ℓ_1}	$N ((w_d - W_d) / (1 - W_d)) (\tilde{Y}_{U_d} (1 - W_d) - (\tilde{\mathbf{x}}_{U_d} - \bar{\mathbf{x}}_U)' \mathbf{B}_{1d})$	$N^2 \left[(w_d^2 / n_d) (1 - f_d) v_{e_{sd}} + (w_d^2 / n_d) (1 - f_d) v_{e_{\bar{s}}} \right]$
\hat{Y}_{d,ℓ_2}	Almost 0	$(N_d^2 (1 - f_d) / n_d) \sum_{s_d} \left((a_{dk} e_{dk} - \overline{a_d e})^2 / (n_d - 1) \right)$
\hat{Y}_{d,ℓ_3}	$N (w_d - W_d) (\tilde{Y}_{U_d} - \tilde{\mathbf{x}}_{U_d}' \mathbf{B}_3)$	$((N_d w_d)^2 (1 - f_d) / n_d) \sum_{s_d} \left((a_{dk} e_{dk} - \overline{a_d e})^2 / (n_d - 1) \right)$
\tilde{Y}_{d,ℓ_1}	$N (w_d - W_d) (\bar{\mathbf{x}}_U - \tilde{\mathbf{x}}_{U_d})' \mathbf{B}_{1d} / (1 - W_d)$	$(N / w_d)^2 \left[(w_d^2 / n_d) (1 - f_d) v_{e_{sd}} + (w_d^2 / n_d) (1 - f_d) v_{e_{\bar{s}}} \right]$
\tilde{Y}_{d,ℓ_2}	Almost 0	$(N_d^2 (1 - f_d) / n_d) \sum_{s_d} \left((a_{dk} e_{dk} - \overline{a_d e})^2 / (n_d - 1) \right)$
\tilde{Y}_{d,ℓ_3}	Almost 0	$(N_d^2 (1 - f_d) / n_d) \sum_{s_d} \left((a_{dk} e_{dk} - \overline{a_d e})^2 / (n_d - 1) \right)$

Table 4
Estimators and Associated Error Terms

Estimator	Ratio Version	Error Term
HT ratio: $\hat{Y}_{d,\ell\ell_1}$	$\hat{Y}_{d,RAT} = \hat{Y}_{d,HT} (X / \hat{X}_{HT})$	$e_{dk} = y_{dk} - \hat{R}_{1d} x_k, \hat{R}_{1d} = \hat{Y}_{d,HT} / \hat{X}_{HT}$
HT post-stratified ratio: $\hat{Y}_{d,\ell r_2}$	$\hat{Y}_{d,POSTR} = \hat{Y}_{d,HT} (X_d / \hat{X}_{d,HT})$	$e_{dk} = y_{dk} - \hat{R}_{2d} x_{dk}, \hat{R}_{2d} = \hat{Y}_{d,HT} / \hat{X}_{d,HT}$
HT alternate ratio: $\hat{Y}_{d,\ell r_3}$	$\hat{Y}_{d,ALTR} = \hat{Y}_{d,HT} + (X_d - \hat{X}_{d,HT}) (\hat{Y}_{HT} / \hat{X}_{HT})$	$e_{dk} = y_{dk} - \hat{R}_3 x_k, \hat{R}_3 = \hat{Y}_{HT} / \hat{X}_{HT}$
Hájek ratio: $\tilde{Y}_{d,\ell\ell_1}$	$\tilde{Y}_{d,RAT} = \hat{Y}_{d,HA} + (X - \hat{X}_{HA}) (\hat{Y}_{d,HT} / \hat{X}_{HT})$	$e_{dk} = y_{dk} - \hat{R}_{1d} x_k, \hat{R}_{1d} = \hat{Y}_{d,HT} / \hat{X}_{HT}$
Hájek alternate ratio: $\tilde{Y}_{d,\ell r_3}$	$\tilde{Y}_{d,ALTR} = \hat{Y}_{d,HA} + (X_d - \hat{X}_{d,HA}) (\hat{Y}_{HT} / \hat{X}_{HT})$	$e_{dk} = y_{dk} - \hat{R}_3 x_k, \hat{R}_3 = \hat{Y}_{HT} / \hat{X}_{HT}$

4.1 Unconditional Results

The unconditional properties of the estimators were assessed using two performance measures: (i) root mean squared error (RMSE) and (ii) coverage rate (CR). They are:

- i. The RMSE is defined as

$$\sqrt{\sum_{m=1}^M (\hat{Y}_d^{(m)} - Y_d)^2 / M},$$

where $\hat{Y}_d^{(m)}$ is the estimated total (either Horvitz-Thompson or Hájek type) based on sample m , and M is the total number of samples drawn for the simulation.

- ii. The coverage rate CR for a given estimator \hat{Y} is defined as the ratio of the number of times that the 95% confidence interval

$$\hat{Y}_d^{(m)} \pm 1.96 \sqrt{v(\hat{Y}_d^{(m)})}$$

contains the true population total to the number of replicates. We used the unconditional variances given by (2.12), and the error terms in Table 4 to estimate the required variances.

The four graphs provided in Figures 1 and 2, summarize the unconditional analysis for small and large domains. Also shown is the impact of increasing $\rho_{x,y}$. The square root of the average mean squared error and coverage rates are used to compare the estimators.

In Figure 1, we note that the RMSE decreases substantially with increasing $\rho_{x,y}$. This can be attributed to the decreasing dispersion of the dependent variable conditional on the independent variable as the correlation between the two increases. We also note that the spread of the RMSE is narrower for the large domain than for the small domain. The ranking of the estimators in terms of RMSE from worst to best is as follows: (i) HT ratio (HT RAT), (ii) Hájek ratio (HA RAT), (iii) HT alternate ratio

(HT ALTR), (iv) Hájek alternate ratio (HA ALTR), and (v) HT post-stratified ratio (HT POSTR). This ranking is in agreement with Result 3.2.

In Figure 2, we note that the unconditional coverage rates are similar across all the estimators regardless of the correlation $\rho_{x,y}$. For small domains the Horvitz-Thompson estimators exhibit a slight degradation in the coverage rate when $\rho_{x,y}$ is weak. But as the correlation increases, their coverage rate becomes comparable to the Hájek type estimators. The Hájek estimators have a better overall coverage rate than their Horvitz-Thompson counterparts.

4.2 Conditional Results

The conditional properties of the estimators were studied using: (i) average relative conditional bias and (ii) conditional coverage rates. They are defined as:

- i. $ARB_d = (100 / M_d) \sum_{m=1}^{M_d} (\hat{Y}_d^{(m)} - Y_d) / Y_d$, where M_d is the number of samples of size n_d .
- ii. The conditional coverage rate has the same definition as its unconditional counterpart. The associated variance is

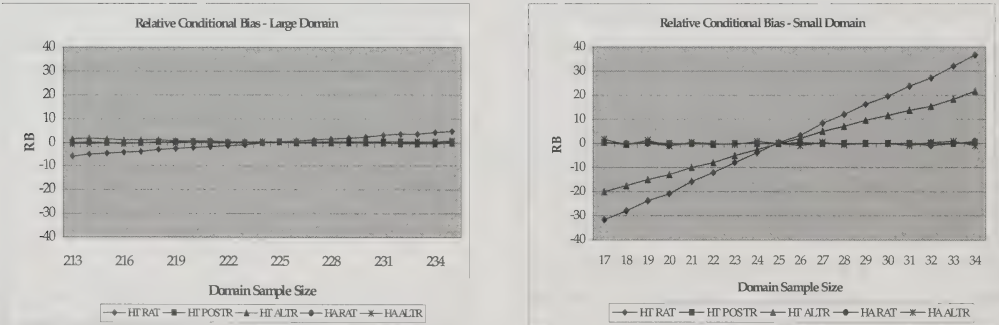
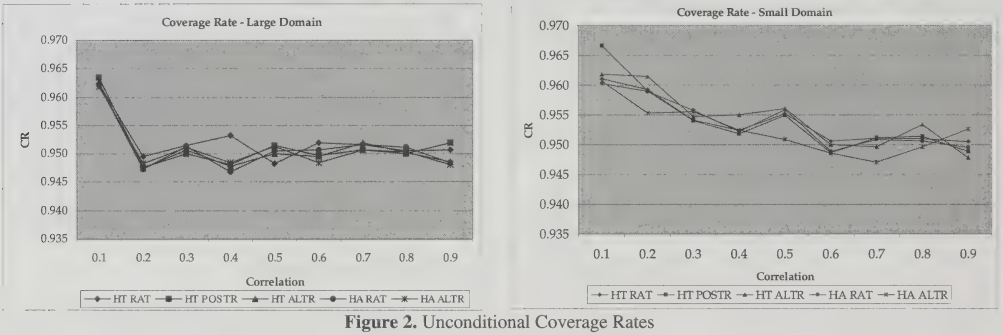
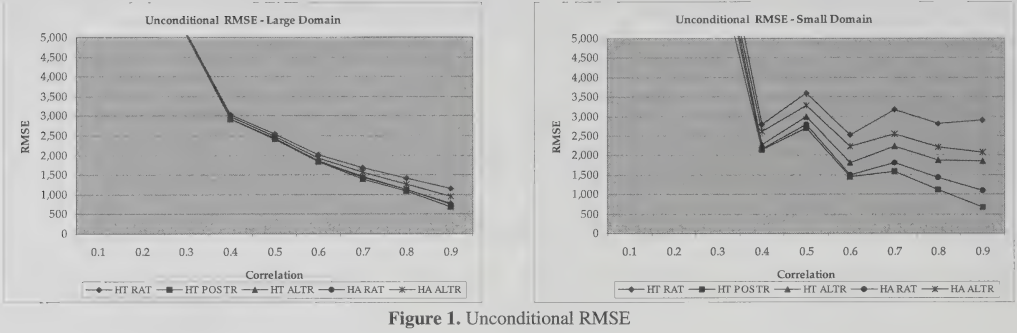
$$v_d^2 = \frac{1}{M_d - 1} \sum_{m=1}^{M_d} (\hat{Y}_d^{(m)} - \bar{Y}_d)^2$$

where

$$\bar{Y}_d = \frac{1}{M_d} \sum_{m=1}^{M_d} \hat{Y}_d^{(m)}.$$

Table 5 summarizes the conditional biases of the ratio versions of estimators (2.1)–(2.4) and (2.6). They were obtained from Table 3 using a single auxiliary variable.

The relative conditional bias and coverage rates of the estimators are summarized in Figures 3, 4a, and 4b with respect to the realized sample size n_d for large and small domains, and for two correlations ($\rho_{x,y} = 0.90$ and $\rho_{x,y} = 0.60$).



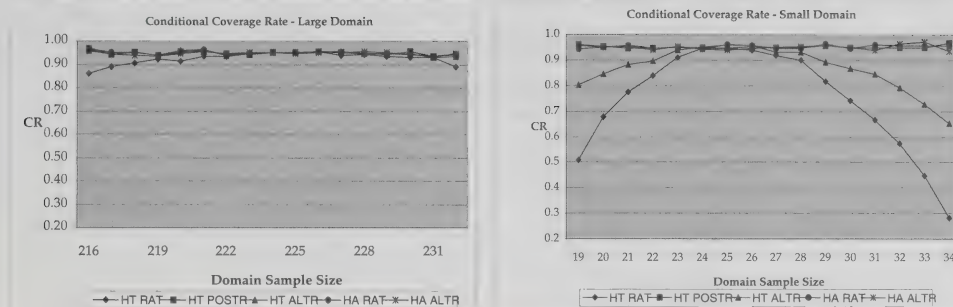


Figure 4a. Conditional Coverage Rates for $\rho_{X,Y} = 0.90$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

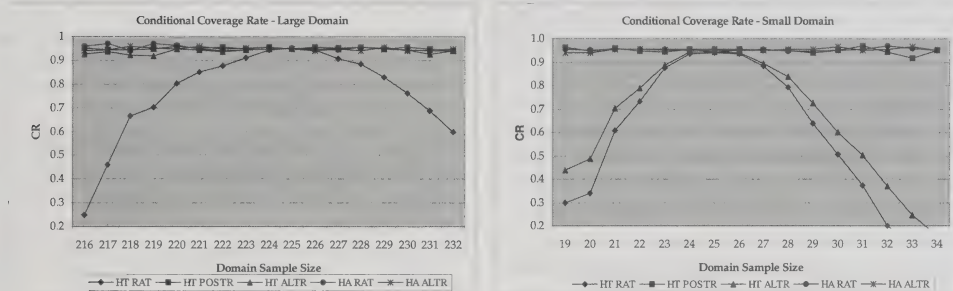


Figure 4b. Conditional Coverage Rates for $\rho_{X,Y} = 0.60$, $\beta_{d1} = 1.0$, and $\beta_{d2} = 3.0$.

Table 5
Conditional Biases of Ratio Versions of Estimators (2.1)-(2.4) and (2.6)

Estimator	Conditional Bias
HT Ratio: \hat{Y}_{d,ℓ_1}	$N \tilde{y}_{U_d} (w_d - W_d) \frac{(\bar{x}_U - W_d \tilde{x}_{U_d})}{\bar{x}_U (1 - W_d)}$
HT post-stratified ratio: \hat{Y}_{d,ℓ_2}	Almost 0
HT alternate ratio: \hat{Y}_{d,ℓ_3}	$N \tilde{y}_{U_d} (w_d - W_d) \frac{(\bar{x}_U - \tilde{x}_{U_d} \bar{y}_U / \tilde{y}_{U_d})}{\bar{x}_U}$
Hájek ratio: \tilde{Y}_{d,ℓ_1}	$N \tilde{y}_{U_d} (w_d - W_d) \frac{W_d (\bar{x}_U - \tilde{x}_{U_d})}{(1 - W_d) \bar{x}_U}$
Hájek alternate ratio: \tilde{Y}_{d,ℓ_3}	Almost 0

The conditional bias presented in Figure 3 supports the theoretical results presented in Table 5. The three Hájek estimators are nearly conditionally unbiased. The magnitude of the conditional bias of both the HT ratio estimator and the HT alternate ratio estimator is in agreement with the theoretical conditional bias. But it should be noted that the conditional bias associated with the HT alternate ratio estimator is smaller than the one of the HT ratio estimator. Also, in larger domains, this conditional bias is less pronounced for the HT alternate ratio estimator.

The conditional coverage rates are given in Figures 4a and 4b. We note that the three Hájek estimators follow closely the nominal 95% coverage probability. The coverage rate of the HT alternate ratio estimator is reasonable in larger domains despite its being conditionally biased. But its coverage deteriorates substantially in smaller domains. The coverage rate of the HT ratio estimator is not acceptable. But it should be noted that the coverage rates of the conditionally biased estimators improve as the realized sample size n_d approaches the expected domain sample size $E(n_d)$.

In summary, the simulation study identified the three Hájek estimators, *Hájek post-stratified ratio*, *Hájek alternate ratio*, and *Hájek ratio* as the best estimators in terms of their conditional and unconditional properties. Note that even though the *Hájek ratio* estimator uses the least domain auxiliary data (it uses domain population counts N_d), its mean squared error is still reasonable. The *Hájek post-stratified ratio* is the best estimator in terms of its conditional and unconditional properties.

5. CONCLUDING REMARKS

We have studied six possible regression estimators of domain totals, each using various levels of auxiliary information at the domain and/or population level. The only estimator that has regression weights that are not domain dependent and that also have the additive property is Horvitz-Thompson estimator $\hat{Y}_{d,\ell\eta}$. This estimator is constructed using auxiliary information at the population level: the domain dependent independent variable y_{dk} is regressed on the auxiliary vector \mathbf{x}_k . However, it can be seriously conditionally biased and the associated confidence intervals can be understated.

The Hájek-type estimators have two the disadvantages: (i) they do not have the additive property; and (ii) their associated regression weights are domain dependent. However, they have the best conditional properties. They are nearly conditionally unbiased, and the conditional confidence intervals associated with the estimators follow closely the nominal coverage rate. They also have the smaller unconditional MSE's. The Hájek estimator that uses the least auxiliary data at the domain level is $\tilde{Y}_{d,\ell\eta}$. It requires domain population counts N_d ($d = 1, \dots, D$), and the population totals \mathbf{X} . Its conditional and unconditional properties are reasonable.

The best Hájek estimator, $\tilde{Y}_{d,\ell\eta}$, uses auxiliary information at the domain level. The Hájek regression type estimator $\tilde{Y}_{d,\ell\eta}$ can be made domain independent using a single set of regression weights as follows. Suppose that the most important domains are $U_g \subseteq U$ ($g = 1, \dots, G$), and that these domains are mutually exclusive and exhaustive. The resulting Hájek estimator is

$$\tilde{Y}_{d,\ell\eta} = \sum_{g=1}^G \left[\hat{Y}_{g,\text{HA}} + (\mathbf{X}_g - \hat{\mathbf{X}}_{g,\text{HA}})' \hat{\mathbf{B}}_{1g} \right]$$

where

$$\hat{Y}_{g,\text{HA}} = (N_g / \hat{N}_g) \hat{Y}_{g,\text{HT}}, \quad \hat{Y}_{g,\text{HT}} = \sum_{s_g} w_k y_{dk}$$

and

$$\hat{\mathbf{B}}_{1g} = \left(\sum_{s_g} w_k \mathbf{x}_k \mathbf{x}_k' / c_k \right)^{-1} \sum_{s_g} w_k \mathbf{x}_k y_{dk} / c_k.$$

REFERENCES

- DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, (Eds. N.L. Johnson and H. Smith). New York: Wiley, Interscience.
- ESTEVAO, V.M., HIDIROGLOU, M.A. and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*. 11, 2, 181-204.
- ESTEVAO, V.M., and SÄRNDAL, C.-E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology*. 213-231.
- HARTLEY, H.O. (1959). *Analytic Studies of Survey Data*. Instituto di Statistica, Rome, Volume in honor of Corrado Gini.
- HIDIROGLOU, M.A. (1991). Structure of the Generalized Estimation System (GES). Statistics Canada report, September, 1991.
- HOLT, D., and SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A*. 142, 33-46.
- RAO, J.N.K. (1985). Conditional Inferences in Survey Sampling. *Survey Methodology*. 15-32.
- SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small Domain Estimation: A conditional analysis. *Journal of American Statistical Association*. 84, 405, 266-275.
- SÄRNDAL, C.-E., SWENSSON B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.

Prediction of Finite Population Totals Based on the Sample Distribution

MICHAEL SVERCHKOV and DANNY PFEFFERMANN¹

ABSTRACT

This article studies the use of the sample distribution for the prediction of finite population totals under single-stage sampling. The proposed predictors employ the sample values of the target study variable, the sampling weights of the sample units and possibly known population values of auxiliary variables. The prediction problem is solved by estimating the expectation of the study values for units outside the sample as a function of the corresponding expectation under the sample distribution and the sampling weights. The prediction mean square error is estimated by a combination of an inverse sampling procedure and a re-sampling method. An interesting outcome of the present analysis is that several familiar estimators in common use are shown to be special cases of the proposed approach, thus providing them a new interpretation. The performance of the new and some old predictors in common use is evaluated and compared by a Monte Carlo simulation study using a real data set.

KEY WORDS: Bootstrap; Design consistency; Informative sampling; Sample-complement distribution.

1. INTRODUCTION

The sample distribution is the parametric distribution of the outcome values for units included in the sample. This distribution is different from the population distribution if the sample selection probabilities are correlated with the values of the study variable even when conditioning on the values of concomitant variables included in the population model. It is also different from the randomization (design) distribution that accounts for all the possible sample selections with the population values held fixed. The sample distribution is defined and discussed with examples in Pfeffermann, Krieger and Rinott (1998), and is further investigated in Pfeffermann and Sverchkov (1999) who use it for the estimation of linear regression models. Krieger and Pfeffermann (1997) use the sample distribution for testing population distribution functions and Pfeffermann and Sverchkov (2003a) discuss its use for fitting Generalized Linear Models. Chambers, Dorfman and Sverchkov (2003) utilize the sample distribution for nonparametric estimation of regression models, and Kim (2002) and Pfeffermann and Sverchkov (2003b) apply it for small area estimation problems.

In this article we study the use of the sample distribution for the prediction of finite population totals under single-stage sampling. It is assumed that the population outcome values (the y -values) are random realizations from some distribution that conditions on known values of auxiliary variables (the x -values). The problem considered is the prediction of the population total Y based on the sample y -values, the sampling weights for units in the sample and the population x -values. The use of the sample distribution

permits conditioning on all these values, which is not possible under the randomization (design) distribution, and the prediction of Y is equivalent therefore to the prediction of the y -values for units outside the sample.

The prediction problem is solved by estimating the conditional expectation of the y -values (given the x -values) for units outside the sample as a function of the conditional sample expectation (the expectation under the sample distribution) and the sampling weights. The prediction mean square error is estimated by a combination of an inverse sampling procedure and a re-sampling method. As it turns out, several familiar estimators in common use and in particular, classical design based estimators are special cases of the proposed procedure, thus providing them a new interpretation. The performance of the new and old predictors is evaluated and compared by mean of a Monte Carlo simulation study using a real data set.

2. THE SAMPLE AND SAMPLE-COMPLEMENT DISTRIBUTIONS

2.1 The Sample Distribution

Suppose that the population values $\{y, X\} = \{(y_1 \dots y_N)', [x_1 \dots x_N]'\}$ are random realizations with conditional probability density function (*pdf*) $f_p(y_i | x_i)$ that may be discrete or continuous. The y -values are assumed to be scalars but the x -values can be vectors. We consider single stage sampling with sample inclusion probabilities $\pi_i = \Pr(i \in s) = g(y, X, Z, i)$ for some function g , where Z defines the population values of design variables used for the sampling process. Note that the y -values are random and we also consider the design variables as random so that the

¹ Michael Sverchkov, The Bureau of Labor Statistics, Washington D.C. 20212, U.S.A.; Danny Pfeffermann, Hebrew University, Israël and University of Southampton, U.K.

g -values are random as well. Let $I_i = 1$ if $i \in s$ and $I_i = 0$, if $i \notin s$. The conditional marginal *sample pdf* is defined as,

$$\begin{aligned} f_s(y_i | \mathbf{x}_i) &\stackrel{\text{def}}{=} f(y_i | \mathbf{x}_i, I_i = 1) \\ &= \frac{\Pr(I_i = 1 | y_i, \mathbf{x}_i) f_p(y_i | \mathbf{x}_i)}{\Pr(I_i = 1 | \mathbf{x}_i)} \end{aligned} \quad (2.1)$$

with the second equality obtained by application of Bayes theorem. Note that $\Pr(I_i = 1 | y_i, \mathbf{x}_i)$ is not necessarily the same as the actual sample selection probability $\pi_i = g(\mathbf{y}, X, Z, i)$ (see Remark 1 below). It follows from (2.1) that the population and sample *pdfs* are different, unless $\Pr(I_i = 1 | y_i, \mathbf{x}_i) = \Pr(I_i = 1 | \mathbf{x}_i)$ for all y_i . When the sample distribution differs from the population distribution it becomes *informative*, and the sampling scheme can not be ignored at the inference process.

Remark 1. It is important to emphasize that the definition and use of the sample distribution does not assume that the sample selection probabilities are function of only (y_i, \mathbf{x}_i) . As mentioned earlier and highlighted by expressing the selection probabilities as $\pi_i = g(\mathbf{y}, X, Z, i)$, the actual selection probabilities may depend on all the population values (\mathbf{y}, X, Z) . However, as shown in Pfeffermann and Sverchkov (1999), $E_p(\pi_i | y_i, \mathbf{x}_i) = \Pr(I_i = 1 | y_i, \mathbf{x}_i)$. Thus, although the selection probabilities may depend on all the population values (\mathbf{y}, X, Z) , for given values (y_i, \mathbf{x}_i) they equal $\Pr(I_i = 1 | y_i, \mathbf{x}_i)$ ‘on average’. In fact, π_i may not depend directly on \mathbf{y} at all and only be a function of (X, Z) , and still the expectation $E_p(\pi_i | y_i, \mathbf{x}_i)$ equals $\Pr(I_i = 1 | y_i, \mathbf{x}_i)$. The reason why the expectation may depend on y_i in this case is that Z may be correlated with y . For example, the 1999 Canadian Workplace and Employee Survey uses a disproportionate stratified sample with the strata defined by region, activity, and the size of the workplace. The size information is obtained from tax records from 1998; see, Patak, Hidioglou and Lavallée (2000) for details. When modeling the payrolls in 1999 against the number of employees, the sampling design is found to be informative, which is explained by the fact that the stratification is based in part on the size obtained from the tax records in the previous year, which are correlated with the payroll the year after. See Fuller (2003) for details of the analysis.

The discussion above should not be understood to mean that π_i is never a function of (y_i, \mathbf{x}_i) only. A classical example for the latter case is retrospective sampling. Thus, in a case control study, the selection probabilities of the cases and controls usually only depend on the respective y and x values (and often just on the y values). In the empirical study of this paper we use a real data set where the sample was drawn by a disproportionate stratified sample

with the strata boundaries defined by the values of the dependent variable.

In what follows we regard the probabilities π_i as random realizations of the random variable $g(\mathbf{y}, X, Z, i)$. Let $w_i = 1/\pi_i$ define the sampling weight of unit i . The following relationships, established in Pfeffermann and Sverchkov (1999) hold for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$, with E_p and E_s defining expectations under the population and sample *pdfs* respectively. (As a special case, $\mathbf{u}_i = y_i$, $\mathbf{v}_i = \mathbf{x}_i$).

$$f_s(\mathbf{u}_i | \mathbf{v}_i) = \frac{E_p(\pi_i | \mathbf{u}_i, \mathbf{v}_i) f_p(\mathbf{u}_i | \mathbf{v}_i)}{E_p(\pi_i | \mathbf{v}_i)} \quad (2.2)$$

$$f_p(\mathbf{u}_i | \mathbf{v}_i) = \frac{E_s(w_i | \mathbf{u}_i, \mathbf{v}_i) f_s(\mathbf{u}_i | \mathbf{v}_i)}{E_s(w_i | \mathbf{v}_i)} \quad (2.3)$$

$$E_p(\mathbf{u}_i | \mathbf{v}_i) = \frac{E_s(w_i \mathbf{u}_i | \mathbf{v}_i)}{E_s(w_i | \mathbf{v}_i)}. \quad (2.4)$$

It follows from (2.4) that

$$\begin{aligned} \text{a)} \quad E_s(w_i | \mathbf{v}_i) &= \frac{1}{E_p(\pi_i | \mathbf{v}_i)}; \\ \text{b)} \quad E_p(\mathbf{u}_i) &= \frac{E_s(w_i \mathbf{u}_i)}{E_s(w_i)}; \\ \text{c)} \quad E_s(w_i) &= \frac{1}{E_p(\pi_i)}. \end{aligned} \quad (2.5)$$

For a detailed discussion of the sample distribution with illustrations, see Pfeffermann *et al.* (1998).

2.2 The Sample-Complement Distribution

Similar to (2.1), we define the conditional *pdf* for units outside the sample as,

$$\begin{aligned} f_c(y_i | \mathbf{x}_i) &\stackrel{\text{def}}{=} f_p(y_i | \mathbf{x}_i, I_i = 0) \\ &= \frac{\Pr(I_i = 0 | y_i, \mathbf{x}_i) f_p(y_i | \mathbf{x}_i)}{\Pr(I_i = 0 | \mathbf{x}_i)}. \end{aligned} \quad (2.6)$$

The relationships (2.2)–(2.5) and the equality $\Pr(I_i = 0 | \mathbf{u}_i, \mathbf{v}_i) = 1 - \Pr(I_i = 1 | \mathbf{u}_i, \mathbf{v}_i) = 1 - E_p(\pi_i | \mathbf{u}_i, \mathbf{v}_i)$ imply the following representations of the sample-complement distribution for general pairs of vector random variables $(\mathbf{u}_i, \mathbf{v}_i)$.

$$\begin{aligned} f_c(\mathbf{u}_i | \mathbf{v}_i) &= \frac{E_p[(1 - \pi_i) | \mathbf{u}_i, \mathbf{v}_i] f_p(\mathbf{u}_i | \mathbf{v}_i)}{E_p[(1 - \pi_i) | \mathbf{v}_i]} \\ &= \frac{E_p[(1 - \pi_i) | \mathbf{u}_i, \mathbf{v}_i]}{E_p[(1 - \pi_i) | \mathbf{v}_i]} \frac{E_p[\pi_i | \mathbf{v}_i]}{E_p[\pi_i | \mathbf{u}_i, \mathbf{v}_i]} f_s(\mathbf{u}_i | \mathbf{v}_i) \end{aligned} \quad (2.7)$$

$$f_c(\mathbf{u}_i | \mathbf{v}_i) = \frac{E_s[(w_i - 1) | \mathbf{u}_i, \mathbf{v}_i] f_s(\mathbf{u}_i | \mathbf{v}_i)}{E_s[(w_i - 1) | \mathbf{v}_i]}. \quad (2.8)$$

(Equation (2.8) follows by application of (2.5a) to the second expression in (2.7)). Also, by (2.8) and the first equation in (2.7),

$$E_c(\mathbf{u}_i | \mathbf{v}_i) = \frac{E_p[(1 - \pi_i) \mathbf{u}_i | \mathbf{v}_i]}{E_p[(1 - \pi_i) | \mathbf{v}_i]} = \frac{E_s[(w_i - 1) \mathbf{u}_i | \mathbf{v}_i]}{E_s[(w_i - 1) | \mathbf{v}_i]}. \quad (2.9)$$

Remark 2. In practical applications the sampling fraction is often very small and hence the sample selection probabilities are small for at least most of the population units. If $\pi_i < \delta$ with probability 1,

$$\begin{aligned} f_c(\mathbf{u}_i | \mathbf{v}_i) &= \frac{E_p[(1 - \pi_i) | \mathbf{u}_i, \mathbf{v}_i] f_p(\mathbf{u}_i | \mathbf{v}_i)}{E_p[(1 - \pi_i) | \mathbf{v}_i]} \\ &= f_p(\mathbf{u}_i | \mathbf{v}_i) + \\ &\quad \frac{E_p\{[E_p(\pi_i | \mathbf{v}_i) - \pi_i | \mathbf{u}_i, \mathbf{v}_i] f_p(\mathbf{u}_i | \mathbf{v}_i)\}}{E_p[(1 - \pi_i) | \mathbf{v}_i]} \\ &= f_p(\mathbf{u}_i | \mathbf{v}_i)(1 + \Delta) \end{aligned} \quad (2.10)$$

where $-\delta < \Delta < \delta/(1 - \delta)$. It follows from (2.10) that for δ sufficiently small, the difference between the population *pdf* and the sample-complement *pdf* is accordingly small, which is not surprising.

3. OPTIMAL PREDICTION OF FINITE POPULATION TOTALS

Let $Y = \sum_{i=1}^N y_i$ define the population total. The problem considered is how to predict Y based on the sample data and possibly population values of auxiliary variables. Denote the 'design information' available for prediction by $D_s = \{(y_i, w_i), i \in s; (\mathbf{x}_j, I_j), j = 1 \dots N\}$ and let $\hat{Y} = \hat{Y}(D_s)$ define the predictor. The MSE of \hat{Y} with respect to the population *pdf* given D_s is,

$$\begin{aligned} \text{MSE}(\hat{Y} | D_s) &= E_p[(\hat{Y} - Y)^2 | D_s] \\ &= E_p\{[\hat{Y} - E_p(Y | D_s)]^2 | D_s\} + V_p(Y | D_s) \\ &= [\hat{Y} - E_p(Y | D_s)]^2 + V_p(Y | D_s) \end{aligned} \quad (3.1)$$

since $[\hat{Y} - E_p(Y | D_s)]$ is fixed given D_s . It follows from (3.1) that $\text{MSE}(\hat{Y} | D_s)$ is minimized when $\hat{Y} = E_p(Y | D_s)$. The latter expectation can be decomposed as,

$$\begin{aligned} E_p(Y | D_s) &= \sum_{i=1}^N E_p(y_i | D_s) \\ &= \sum_{i \in s} E_p(y_i | D_s, I_i = 1) + \sum_{j \notin s} E_p(y_j | D_s, I_j = 0) \\ &= \sum_{i \in s} y_i + \sum_{j \notin s} E_c(y_j | D_s) \\ &= \sum_{i \in s} y_i + \sum_{j \notin s} E_c(y_j | \mathbf{x}_j) \end{aligned} \quad (3.2)$$

where in the last equality we assume that y_j for $j \notin s$ and D_s are uncorrelated given \mathbf{x}_j . The prediction problem reduces therefore to the estimation of the expectations $E_c(y_j | \mathbf{x}_j)$. In section 4 we consider semi-parametric estimation of these expectations.

4. SEMI-PARAMETRIC PREDICTION OF FINITE POPULATION TOTALS

Suppose that the sample-complement model takes the form,

$$\begin{aligned} y_j &= C_\beta(\mathbf{x}_j) + \varepsilon_j, \\ E_c(\varepsilon_j | \mathbf{x}_j) &= 0, E_c(\varepsilon_j^2 | \mathbf{x}_j) = \sigma^2 v(\mathbf{x}_j), \\ E_c(\varepsilon_k \varepsilon_j | \mathbf{x}_k, \mathbf{x}_j) &= 0, k \neq j \end{aligned} \quad (4.1)$$

where $C_\beta(\mathbf{x})$ is a known (possibly nonlinear) function of \mathbf{x} that depends on an unknown vector parameter β . The variances $\sigma^2 v(\mathbf{x}_j)$ are assumed known except for σ^2 .

Remark 3. In actual applications the model (4.1) can be identified by a two-step procedure, utilizing the equality $E_c(y_i | \mathbf{x}_i) = E_s(r_i y_i | \mathbf{x}_i)$ with $r_i = (w_i - 1) / E_s[(w_i - 1) | \mathbf{x}_i]$ (follows from Equation 2.9). First, estimate $E_s(w_i | \mathbf{x}_i)$ and hence r_i by regressing w_i against \mathbf{x}_i using the sample data. Let $\hat{r}_i = (w_i - 1) / [\hat{E}_s(w_i | \mathbf{x}_i) - 1]$ and transform $y_i^* = \hat{r}_i y_i$. Second, study the relationship in the sample between y_i^* and \mathbf{x}_i for identifying the form of $C_\beta(\mathbf{x}_i)$. See Pfeiffermann and Sverchkov (1999, 2003a) for examples of estimating $E_s(w_i | \mathbf{x}_i)$. A similar procedure can be applied for identifying the variance function $v(\mathbf{x}_i)$, using the empirical residuals $\hat{\varepsilon}_i = y_i - \hat{E}_s(\hat{r}_i y_i | \mathbf{x}_i)$.

The function $C_\beta(\mathbf{x}_j)$ in (4.1) with the true vector parameter β satisfies for all $j \notin s$,

$$\begin{aligned} C_\beta(\mathbf{x}_j) &= \arg \min_{C_\beta(\mathbf{x}_j)} E_c \left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \middle| \mathbf{x}_j \right) \\ &= \arg \min_{C_\beta(\mathbf{x}_j)} E_s \left(r_i \frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right). \end{aligned} \quad (4.2)$$

(The second equality follows from (2.9)). Hence, by substituting the sample expectation outside the curved brackets by the sample mean (a straightforward application

of the method of moments) and estimating r_i by \hat{r}_i (see Remark 3), the vector β can be estimated as,

$$\hat{\beta}_1 = \arg \min_{\beta} \sum_{i \in s} \left(\hat{r}_i \frac{[y_i - C_{\beta}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)} \right). \quad (4.3)$$

The predictor of the population total takes then the form,

$$\hat{Y}_1 = \sum_{i \in s} y_i + \sum_{j \in s} C_{\hat{\beta}_1}(\mathbf{x}_j). \quad (4.4)$$

Alternatively, it follows from (4.1) that,

$$\begin{aligned} E_c \left(\frac{[y_j - C_{\beta}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \middle| \mathbf{x}_j \right) \\ = E_c \left(\frac{[y_j - C_{\beta}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right) \\ = E_s \left(\left[\frac{w_j - 1}{E_s(w_j) - 1} \right] \frac{[y_j - C_{\beta}(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right) \end{aligned} \quad (4.5)$$

where the right hand side expectation is with respect to the joint distribution of (y_i, \mathbf{x}_j) . Thus, β can be estimated as,

$$\hat{\beta}_2 = \arg \min_{\beta} \sum_{i \in s} (w_i - 1) \frac{[y_i - C_{\beta}(\mathbf{x}_i)]^2}{v(\mathbf{x}_i)} \quad (4.6)$$

since $E_s(w_i) = \text{constant}$. The predictor of Y with β estimated by $\hat{\beta}_2$ is therefore,

$$\hat{Y}_2 = \sum_{i \in s} y_i + \sum_{j \in s} C_{\hat{\beta}_2}(\mathbf{x}_j). \quad (4.7)$$

Remark 4. A notable advantage of the use of the predictor \hat{Y}_2 over the use of the predictor \hat{Y}_1 is that it does not require the identification and estimation of the expectation $w(\mathbf{x}) = E_s(w | \mathbf{x})$. On the other hand, in situations where this expectation can be estimated properly, the predictor \hat{Y}_1 is likely to be more accurate since the weights $r_i = (w_i - 1) / [E_s(w | \mathbf{x}_i) - 1]$ will often be less variable than the weights $(w_i - 1)$. This is because the weights r_i only account for the net effect of the sampling process on the target conditional distribution $f_c(y_i | \mathbf{x}_i)$, whereas the weights $(w_i - 1)$ account for the effect of the sampling process on the joint distribution $f_c(y_i, \mathbf{x}_i)$. In particular, when w_i is a deterministic function of \mathbf{x}_i such that $w_i = w(\mathbf{x}_i)$, the sampling process is noninformative and $f_c(y_i | \mathbf{x}_i) = f_s(y_i | \mathbf{x}_i) = f_p(y_i | \mathbf{x}_i)$. In this case the estimator $\hat{\beta}_1$ (but not $\hat{\beta}_2$) coincides with the optimal generalized least square (GLS) estimator of β since $r_i = 1$ and the model (4.1) holds for the sample data. (For the data analysed in section 7, the empirical variance of the weights

r_i is 1.36, whereas the empirical variance of the weights w_i is 2.66). In contrast to this, when the sampling weights w_i are independent of \mathbf{x}_i , the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, and hence the predictors \hat{Y}_1 and \hat{Y}_2 are equal since $w(\mathbf{x}_i) = \text{constant}$.

An interesting special case of the predictor \hat{Y}_2 arises when the working model postulated for the sample-complement is linear with an intercept term and constant variance. Let $\mathbf{x}'_i = (1, \tilde{\mathbf{x}}'_i)$. As easily verified, the estimator in this case takes the form,

$$\hat{Y}_{2, \text{Reg}} = \sum_{i \in s} y_i + \hat{Y}_c + \tilde{B}'_c [\tilde{X}(c) - \hat{X}_c] \quad (4.8)$$

where $\tilde{X}(c) = \sum_{i \in s} \tilde{\mathbf{x}}_i$, $(\hat{Y}_c, \hat{X}_c) = [(N - n) / \sum_{i \in s} (w_i - 1)] [\sum_{i \in s} (w_i - 1)(y_i, \tilde{\mathbf{x}}_i)]$ and \tilde{B}_c is the probability weighted estimator of the vector coefficient of $\tilde{\mathbf{x}}_i$ but with the weights $(w_i - 1)$ instead of w_i .

Remark 5. The predictor $\hat{Y}_{2, \text{Reg}}$ can be obtained as a special case of the Cosmetic predictors proposed by Brewer (1999). It should be emphasized, however, that the development of the cosmetic predictors and the derivation of their MSE assumes explicitly *noninformative* sampling.

An important property of $\hat{Y}_{2, \text{Reg}}$ is that under general conditions it is *design consistent* for Y , irrespective of the true sample-complement model (see Lemma 1 below). Many analysts view ‘design consistency’ as an essential requirement from any predictor; see the discussion in Hansen, Madow and Tepping (1983) and Samdal (1980). The following Lemma 1 defines conditions under which the more general predictor \hat{Y}_2 of (4.7) is design consistent for Y .

Lemma 1. The predictor \hat{Y}_2 is design consistent for Y if the working model used for the computation of $\hat{\beta}_2$ satisfies the conditions, *i*- $C_{\beta}(\mathbf{x})$ has an intercept term, *ii*- $C_{\beta}(\mathbf{x})$ is differentiable with respect to β in the neighborhood of $\hat{\beta}_2$ and *iii*- $v(\mathbf{x}) = \text{constant}$.

Proof: By (4.6) and condition *iii*, $\hat{\beta}_2 = \arg \min_{\beta} \sum_{i \in s} (w_i - 1) [y_i - C_{\beta}(\mathbf{x}_i)]^2$ and by condition *i*, $C_{\beta}(\mathbf{x}) = \beta_0 + C_{\beta_1, \beta_p}(\tilde{\mathbf{x}})$, so that by condition *ii*, $\partial / \partial \beta_0 \{ \sum_{i \in s} (w_i - 1) [y_i - C_{\beta}(\mathbf{x}_i)]^2 \}_{\beta = \hat{\beta}_2} = 0$, which implies $\sum_{i \in s} (w_i - 1) [y_i - C_{\hat{\beta}_2}(\mathbf{x}_i)] = 0$ or,

$$\sum_{i \in s} w_i y_i = \sum_{i \in s} y_i + \sum_{i \in s} w_i C_{\hat{\beta}_2}(\mathbf{x}_i) - \sum_{i \in s} C_{\hat{\beta}_2}(\mathbf{x}_i). \quad (4.9)$$

The proof is completed by noting that under mild regularity conditions $\sum_{i \in s} w_i y_i$ is design consistent for Y , and $\sum_{i \in s} w_i C_{\hat{\beta}_2}(\mathbf{x}_i)$ is design consistent for $\sum_{j=1}^N C_{\hat{\beta}_2}(\mathbf{x}_j)$. Thus, the right hand side of (4.9) converges in probability to \hat{Y}_2 while the left hand side converges in probability to Y .

It is important to emphasize again that the Lemma does not assume that the working model is the correct sample-complement model.

The use of the predictors \hat{Y}_1 and \hat{Y}_2 requires a specification of the sample-complement model. Next we develop another predictor that only requires the identification and estimation of the sample model. The approach leading to this predictor is a sample-complement analogue of the 'bias correction method' proposed by Chambers *et al.* (2003). The proposed predictor is based on the following relationship,

$$\begin{aligned} \sum_{j \in s} E_c(y_j | \mathbf{x}_j) &= \sum_{j \in s} E_s(y_j | \mathbf{x}_j) \\ &+ (N-n) \left\{ \frac{1}{N-n} \sum_{j \in s} E_c \left\{ [y_j - E_s(y_j | \mathbf{x}_j)] | \mathbf{x}_j \right\} \right\} \\ &\equiv \sum_{j \in s} E_s(y_j | \mathbf{x}_j) \\ &+ (N-n) \left\{ \frac{1}{N-n} \sum_{j \in s} E_c [y_j - E_s(y_j | \mathbf{x}_j)] \right\} \end{aligned} \quad (4.10)$$

where in the second row we replaced the sample-complement average of the conditional expectations $E_c(y_j | \mathbf{x}_j)$ by its expectation over the sample-complement distribution of the \mathbf{x} -values (n denotes the sample size). By (2.9),

$$\begin{aligned} E_c[y_j - E_s(y_j | \mathbf{x}_j)] \\ = E_s \left\{ \frac{w_j - 1}{[E_s(w_j) - 1]} [y_j - E_s(y_j | \mathbf{x}_j)] \right\} \end{aligned} \quad (4.11)$$

implying that the sample-complement mean in the second row of (4.10) can be estimated as $\hat{M}_c = 1/n \sum_{i \in s} \{[(w_i - 1)/(\bar{w}_s - 1)][y_i - \hat{E}_s(y_i | \mathbf{x}_i)]\}$, where $\bar{w}_s = \sum_{i \in s} w_i / n$. The proposed predictor therefore takes the form,

$$\hat{Y}_3 = \sum_{i \in s} y_i + \sum_{j \in s} \hat{E}_s(y_j | \mathbf{x}_j) + (N-n) \hat{M}_c \quad (4.12)$$

with $\hat{E}_s(y_j | \mathbf{x}_j)$ estimated from the sample data. The use of \hat{Y}_3 only requires the identification and estimation of the sample regression $E_s(y_j | \mathbf{x}_j)$, which can be carried out using conventional regression techniques. Moreover, under mild conditions \hat{Y}_3 is *design consistent* for Y even if the expectation $E_s(y_j | \mathbf{x}_j)$ is misspecified. This property follows from the fact that $\sum_{j \in s} \hat{E}_s(y_j | \mathbf{x}_j)$ is design consistent for $\sum_{j \in s} E_s(y_j | \mathbf{x}_j)$ and $(N-n) \hat{M}_c$ is design consistent for $M_c = \sum_{j \in s} [y_j - E_s(y_j | \mathbf{x}_j)]$.

Remark 6. If the model fitted to the sample data is linear regression with an intercept and constant residual variance, the difference between the predictor $\hat{Y}_{2, \text{Reg}}$ defined by (4.8) and the predictor \hat{Y}_3 is that $\hat{Y}_{2, \text{Reg}}$ uses a consistent estimator for the regression coefficients defining the linear

approximation to the model holding for the sample-complement, whereas in \hat{Y}_3 the regression coefficients are estimated by ordinary least squares (OLS), thus estimating the linear approximation to the sample model.

Finally, rather than only predicting the sample-complement values as with the previous predictors, one could instead predict all the population values by their estimated expectations under the population model. Assuming that the latter model is linear regression with an intercept term and constant residual variance, application of (2.5b) yields,

$$\begin{aligned} \beta &= \arg \min_{\beta} E_p(y_k - \mathbf{x}_k' \beta)^2 \\ &= \arg \min_{\beta} \frac{E_s[w_k (y_k - \mathbf{x}_k' \beta)^2]}{E_s(w_k)}. \end{aligned} \quad (4.13)$$

Estimating the sample expectation in the numerator of (4.13) by the corresponding sample mean (application of the method of moments) and minimizing the sample mean with respect to β yields the familiar probability weighted estimator $\hat{B}_{pw} = (X'_{[s]} W_s X_{[s]})^{-1} (X'_{[s]} W_s Y_s)$, where $(X_{[s]}, Y_s) = \{(\mathbf{x}_1 \dots \mathbf{x}_n)', (y_1 \dots y_n)'\}$ and $W_s = \text{Diag}[w_1 \dots w_n]$. Let $\mathbf{x}'_i = (1, \tilde{\mathbf{x}}'_i)$. Estimating $\hat{E}_p(y_k | \mathbf{x}_k) = \mathbf{x}'_k \hat{B}_{pw} = \hat{B}_0 + \tilde{\mathbf{x}}'_k \hat{B}_{pw}$ and summing over all the population values yields the familiar generalized regression (GREG) estimator (Särndal 1980),

$$\begin{aligned} \hat{Y}_{\text{GREG}} &= N \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i} + \tilde{B}'_{pw} \left[\tilde{X}(p) - N \frac{\sum_{i \in s} w_i \mathbf{x}_i}{\sum_{i \in s} w_i} \right]; \\ \tilde{X}(p) &= \sum_{k=1}^N \tilde{\mathbf{x}}_k. \end{aligned} \quad (4.14)$$

Remark 7. By considering the estimation of Y as a prediction problem, the use of the predictor $\hat{Y}_{2, \text{Reg}}$ in (4.8) requires the prediction of $(N-n)$ values whereas the use of the GREG requires the prediction of N values. Hence, in situations where both the sample-complement model and the population model can be approximated fairly well by linear regression models with intercept terms (but possibly with different vectors of coefficients for the two models), one expects that for sufficiently large sampling fractions n/N the predictor $\hat{Y}_{2, \text{Reg}}$ will be superior (see the empirical results in section 7).

5. EXAMPLES

5.1 Prediction with No Concomitant Variables

Let $\mathbf{x}_i = 1$ for all i . By (3.2),

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{j \notin s} \hat{E}_c(y_j) = \sum_{i \in s} y_i + (N-n) \hat{E}_s \left(\frac{w_j - 1}{\hat{E}_s(w_j) - 1} y_j \right). \quad (5.1)$$

Estimating the two sample expectations in the right hand side of (5.1) by the respective sample means yields the estimator,

$$\begin{aligned} \hat{Y}_{El} &= \sum_{i \in s} y_i + (N-n) \frac{1}{n} \sum_{i \in s} \frac{w_i - 1}{\bar{w}_s - 1} y_i \\ &= \sum_{i \in s} y_i + \frac{(N-n)}{\sum_{i \in s} (w_i - 1)} \sum_{i \in s} (w_i - 1) y_i. \end{aligned} \quad (5.2)$$

In (5.2), $\sum_{i \in s} (w_i - 1) y_i$ is a 'Horvitz-Thompson estimator' of $\sum_{j \in s} y_j$. The multiplier $(N-n) / \sum_{i \in s} (w_i - 1)$ is a 'Hajek type correction' for controlling the variability of the sampling weights. Notice that \hat{Y}_{El} is a special case of the predictor $\hat{Y}_{2, \text{Reg}}$ defined in (4.8), obtained by setting $\mathbf{x}_i = 1$ for all i . It is also a special case of the predictor \hat{Y}_3 if one estimates $\hat{E}_s(y_j) = \bar{y} = \sum_{i \in s} y_i / n$. For sampling designs such that $\sum_{i \in s} w_i = N$ for all s , or if one estimates $\hat{E}_s(w_i) = N/n$, the predictor \hat{Y}_{El} reduces to the familiar Horvitz-Thompson estimator of the population total, $\hat{Y}_{H-T} = \sum_{i \in s} w_i y_i$.

As with the GREG estimator considered in section 4, rather than predicting the sample-complement total $Y_c = \sum_{j \notin s} y_j$ and using the predictor \hat{Y}_{El} , one could predict all the population y -values by estimating their expectations under the population model. By (2.5b), $E_p(y_i) = E_s(w_i y_i) / E_s(w_i)$. Estimating the two sample expectations by the corresponding sample means yields the familiar Hajek estimator,

$$\begin{aligned} \hat{Y}_{\text{Hajek}} &= \sum_{k=1}^N \hat{E}_p(y_k) = N \hat{E}_s \left(\frac{w_i y_i}{\hat{E}_s(w_i)} \right) \\ &= \frac{N}{\sum_{i \in s} w_i} \sum_{i \in s} w_i y_i. \end{aligned} \quad (5.3)$$

Here again, we anticipate \hat{Y}_{El} to be more precise than \hat{Y}_{Hajek} as the sampling fraction increases (see also the empirical results in section 7). Note that \hat{Y}_{El} and \hat{Y}_{Hajek} are the same and coincide with the Horvitz-Thompson estimator for sampling designs satisfying $\sum_{i \in s} w_i = N$.

5.2 Optimal Prediction with Concomitant Variables, Comparison with Optimal Predictors Under Noninformative Sampling

Let the population model be,

$$\begin{aligned} y_i &= H_\beta(\mathbf{x}_i) + \varepsilon_i, \quad E_p(\varepsilon_i | \mathbf{x}_i) = 0, \\ E_p(\varepsilon_i^2 | \mathbf{x}_i) &= v(\mathbf{x}_i), \quad E_p(\varepsilon_i \varepsilon_j | \mathbf{x}_i, \mathbf{x}_j) = 0, \quad i \neq j \end{aligned} \quad (5.4)$$

and suppose that the sample inclusion probabilities can be modeled as,

$$\pi_i = K \times [y_i g(\mathbf{x}_i) + \delta_i], \quad E_p(\delta_i | \mathbf{x}_i, y_i) = 0 \quad (5.5)$$

where $H_\beta(\mathbf{x})$, $v(\mathbf{x})$ and $g(\mathbf{x})$ are positive functions and K is a normalizing constant. (Below we consider the special case of 'regression through the origin'). This sampling scheme is considered for illustration only, although in section 2 we mention several practical situations where the sample selection probabilities depend directly on the y and \mathbf{x} -values. In particular, this is the case with the data set analysed in section 7. Under (5.4) and (5.5), $\pi(\mathbf{x}_i) = E_p(\pi_i | \mathbf{x}_i) = K H_\beta(\mathbf{x}_i) g(\mathbf{x}_i)$. Hence, by (2.9), (5.4) and (5.5),

$$\begin{aligned} E_c(y_j | \mathbf{x}_j) &= E_p \left(\frac{1 - \pi_j}{1 - \pi(\mathbf{x}_j)} y_j | \mathbf{x}_j \right) \\ &= E_p \left(\frac{1 - \pi(\mathbf{x}_j) - K \varepsilon_j g(\mathbf{x}_j) - K \delta_j}{1 - \pi(\mathbf{x}_j)} y_j | \mathbf{x}_j \right) \\ &= E_p(y_j | \mathbf{x}_j) - \frac{K g(\mathbf{x}_j) v(\mathbf{x}_j)}{1 - \pi(\mathbf{x}_j)}. \end{aligned} \quad (5.6)$$

The last expression in (5.6) shows that $E_c(y_j | \mathbf{x}_j) < E_p(y_j | \mathbf{x}_j) = H_\beta(\mathbf{x}_j)$, which is clear since for the inclusion probabilities defined by (5.5), the sample-complement tends to include the units with the smaller y -values for any given \mathbf{x} -values. Note, however, that as $n/N \rightarrow 0$, $K \rightarrow 0$ and $E_p(y_j | \mathbf{x}_j) - E_c(y_j | \mathbf{x}_j) \rightarrow 0$ (see Remark 2).

As a special case of (5.4), consider the case of a single auxiliary variable x and let $H_\beta(x) = x\beta$ and $v(x) = \sigma^2 x$ ('regression through the origin with variance proportional to x '). For *noninformative* sampling and known β , the optimal unbiased predictor of Y minimizing $E_p[(\hat{Y} - Y)^2 | D_s]$ is in this case, $\hat{Y} = \sum_{i \in s} y_i + \beta \sum_{j \notin s} x_j$. In the practical case of unknown β , the optimal unbiased predictor of Y is the familiar Ratio estimator $\hat{Y}_R = N \bar{y}(\bar{X} / \bar{x})$ with \bar{y} denoting the sample mean of Y and (\bar{x}, \bar{X}) denoting the sample and population means of x (Brewer 1963, Royall 1970).

Now let $g(x) = 1$ in (5.5) for all x , so that $\pi_i = n(y_i + \delta_i) / \sum_{j=1}^N (y_j + \delta_j)$. For sufficiently large N , we can approximate $\pi_i \approx n(y_i + \delta_i) / (N \beta \bar{X})$, implying that $\pi(x_i) = E_p(\pi_i | x_i) \approx n x_i / (N \bar{X})$. By (5.6), $E_c(y_j | x_j) = x_j \beta - \sigma^2 x_j / [\beta(f^{-1} \bar{X} - x_j)]$ where $f = n/N$ is the sampling fraction, so that for known β and σ^2 the optimal predictor of Y is,

$$\hat{Y}_{E, \text{Reg}} = \sum_{i \in s} y_i + \beta \sum_{j \notin s} x_j - \frac{\sigma^2}{\beta} \sum_{j \notin s} \frac{x_j}{f^{-1} \bar{X} - x_j}. \quad (5.7)$$

Lemma 2: Let the population model be defined by (5.4) with $H_\beta(x) = x\beta$ and $v(x) = \sigma^2 x$. Assume also

$E_p(\epsilon_i^3 | x_i) = 0$. Suppose that the sample units are selected independently with probabilities defined as in (5.5), with $g(x) = 1$. Then,

$$\text{MSE}_p(\hat{Y}_{E, \text{Reg}} | D_s) = \sigma^2 \sum_{j \in s} x_j - (\sigma^2 / \beta)^2 \sum_{j \in s} [x_j^2 / (f^{-1} \bar{X} - x_j)^2]. \quad (5.8)$$

Proof: By the independence of the population values and of the sample selections,

$$\begin{aligned} \text{MSE}_p(\hat{Y}_{E, \text{Reg}} | D_s) &= E_p[(\hat{Y}_{E, \text{Reg}} - Y)^2 | D_s] \\ &= \sum_{j \in s} E_c\{[y_j - E_c(y_j | x_j)]^2 | x_j\}. \end{aligned}$$

By (5.6), $[y_j - E_c(y_j | x_j)]^2 = \{\epsilon_j + x_j^* / [1 - \pi(x_j)]\}^2$ where $x_j^* = K\sigma^2 x_j$, $K = n / \beta N \bar{X}$ and $\pi(x_j) = E_p(\pi_j | x_j) = nx_j / (N \bar{X})$. Hence,

$$\begin{aligned} E_c\{[y_j - E_c(y_j | x_j)]^2 | x_j\} &= E_c(\epsilon_j^2 | x_j) + 2x_j^* / (1 - \pi(x_j)) E_c(\epsilon_j | x_j) \\ &\quad + [x_j^* / (1 - \pi(x_j))]^2. \end{aligned}$$

Now,

$$\begin{aligned} E_c(\epsilon_j^2 | x_j) &= E_p[1 - \pi_j / (1 - \pi(x_j)) \epsilon_j^2 | x_j] \\ &= E_p[1 - \pi(x_j) - K\epsilon_j - K\delta_j / (1 - \pi(x_j)) \epsilon_j^2 | x_j] \\ &= E_p(\epsilon_j^2 | x_j) = \sigma^2 x_j \end{aligned}$$

and

$$\begin{aligned} E_c(\epsilon_j | x_j) &= E_p[1 - \pi(x_j) - K\epsilon_j - K\delta_j / (1 - \pi(x_j)) \epsilon_j | x_j] \\ &= -x_j^* / (1 - \pi(x_j)). \end{aligned}$$

It follows therefore that $\text{MSE}_p(\hat{Y}_{E, \text{Reg}} | D_s) = \sigma^2 \sum_{j \in s} x_j - \sum_{j \in s} [x_j^* / (1 - \pi(x_j))]^2$. Q.E.D.

Remark 8: For noninformative sampling and with known β , the prediction MSE of the optimal predictor $\hat{Y} = \sum_{i \in s} y_i + \beta \sum_{j \in s} x_j$ is, $E_p[(\hat{Y} - Y)^2 | D_s] = \sigma^2 \sum_{j \in s} x_j$. This MSE is larger than the MSE obtained under the informative sampling scheme defined by the Lemma, which is obvious since the latter scheme tends to sample the units with the larger y-values and hence also with the larger x-values and the larger standard deviations.

6. MEAN SQUARE ERROR ESTIMATION

Estimating $\text{MSE}(\hat{Y} | D_s) = E_p[(\hat{Y} - Y)^2 | D_s]$ for the predictors \hat{Y} considered in section 4 requires strict model assumptions that could be hard to validate. This is largely

due to the conditioning on the design information D_s . In order to deal with this problem, we propose to estimate instead the unconditional MSE, $\text{MSE}(\hat{Y}) = E[(\hat{Y} - Y)^2] = E_{D_s}\{E_p[(\hat{Y} - Y)^2 | D_s]\}$, where $E_{D_s} = E_D E_s$ defines the expectation over the sample distribution (given the selected sample) and over all possible sample selections. Notice that $E_p[(\hat{Y} - Y)^2 | D_s]$ can be viewed as a random variable $u(D_s)$, so that $\text{MSE}(\hat{Y}) = E_{D_s}[u(D_s)]$ defines its 'best predictor' with respect to the mean square loss function under the distribution f_{D_s} over which the expectation E_{D_s} is taken. By changing the order of the expectations, the unconditional MSE can be expressed as,

$$\begin{aligned} \text{MSE}(\hat{Y}) &= E_s E_p E_D[(\hat{Y} - Y)^2 | y] \\ &= E_p E_D[(\hat{Y} - Y)^2 | y] \end{aligned} \quad (6.1)$$

where $y = \{y_i; i \in U\}$. Estimating the unconditional MSE of any of the predictors \hat{Y} can be carried out therefore by estimating its randomization MSE, see Pfeffermann (1993) for further discussion. Estimation of the randomization MSE of the various predictors has the additional advantage of allowing their use under the design based approach.

Estimation of randomization variances of design based estimators is considered extensively in the literature and many diverse methods are in routine use. However, in view of the complicated structure of some of the predictors considered in this study and in order not to restrict to particular sampling schemes, we propose below the use of a two-step procedure that combines an inverse sampling process (Step 1) and what can be viewed as a bootstrap resampling algorithm (Step 2). A notable advantage of this procedure is that it is general and applies 'equally' to all the predictors. Also, unlike other variance estimation methods in common use, it does not require knowledge of the pair wise joint selection probabilities $\pi_{ij} = \Pr(i, j \in s)$. As discussed later, a valid application of the first step requires sufficiently large samples. The two steps of the proposed procedure are as follows:

Step 1- Generate a single 'pseudo population' by selecting *with replacement* N units from the original sample with probabilities proportional to $w_i = 1/\pi_i$, where N is the population size. The justification for this step is given below, see also Remark 10. Denote by Y_{pp} the sum of the y-values in the pseudo population.

Step 2- Select independently a large number B of bootstrap samples from the pseudo population generated in Step 1, using the same sampling scheme as used for the selection of the original sample, and re-estimate the population total.

Let \hat{Y} represent any of the predictors and denote the predictor obtained for bootstrap sample b by \hat{Y}_{pp}^b . Estimate,

$$\hat{E}_D(\hat{Y} - Y)^2 = \frac{1}{B} \sum_{b=1}^B (\hat{Y}_{pp}^b - Y_{pp})^2. \quad (6.2)$$

The performance of the estimator (6.2) in estimating the randomization MSE depends obviously on the ‘closeness’ of the pseudo population generated in Step 1 to the actual population from which the original sample was drawn. The closeness of the two populations can be verified in part by noting that the marginal distribution of $y_i | \mathbf{x}_i$ in the pseudo population is the same as in the original population. To see this, note that the pseudo population generated in Step 1 is a ‘sample with replacement’ from the original sample with selection probabilities Cw_i on each draw, where $C = 1/\sum_{i=1}^n w_i$. Denoting by $f_{pp}(y_i | \mathbf{x}_i)$ the marginal pseudo population distribution we find using (2.2) and (2.5a),

$$\begin{aligned} f_{pp}(y_i | \mathbf{x}_i) &= \frac{E_s(Cw_i | y_i, \mathbf{x}_i) f_s(y_i | \mathbf{x}_i)}{E_s(Cw_i | \mathbf{x}_i)} \\ &= \frac{E_p(\pi_i | \mathbf{x}_i) f_s(y_i | \mathbf{x}_i)}{E_p(\pi_i | y_i, \mathbf{x}_i)} = f_p(y_i | \mathbf{x}_i). \end{aligned} \quad (6.3)$$

Remark 9. Equation (6.3) only refers to the marginal distribution of $y_i | \mathbf{x}_i$. Like with the standard bootstrap method, a successful application of the proposed procedure requires that the original sample size is sufficiently large and that the sample measurements are approximately independent. Pfeffermann *et al.* (1998) establish conditions under which for independent population measurements the sample measurement are ‘asymptotically independent’ under commonly used sampling schemes with unequal selection probabilities.

Remark 10. Step 1 is similar and asymptotically equivalent to duplicating sample unit i w_i times. Notice, however, that the use of this duplication procedure does not yield pseudo populations of size N unless $\sum_{i=1}^n w_i = N$. It is also not clear how to establish the relationship (6.3) when using this procedure.

7. EMPIRICAL ILLUSTRATIONS

7.1 Description of Empirical Study

In order to illustrate the performance of the predictors and the associated MSE estimates discussed in previous sections we use a real data set, collected as part of the 1988 U.S. National Maternal and Infant Health Survey. The survey uses a disproportionate stratified random sample of vital records with the strata defined by *mother’s race* and *child’s birth weight*; see Korn and Graubard (1995) for details. For the empirical study in this section we considered the sample data as ‘population’ and selected independently

1,000 samples with probabilities proportional to the inverse of the original sampling weights, using a systematic PPS sampling scheme. The list of ‘population units’ was randomly ordered before every sample selection. For each sample we predicted the population total of *birth weight* (measured in *grams*, divided by 10,000 in the present study), using *gestational age* as the auxiliary variable (measured in *weeks*). The sample inclusion probabilities depend therefore on the values of the study variable that defines the original strata. Notice that although the original sample was supposedly a stratified random sample, the sampling weights actually vary within the strata, which is why we used systematic PPS sampling for the simulation study. We considered three different sample sizes, $n = 232, 1,145, 2,429$. The ‘population’ (original sample) size is $N = 9,948$. (For $n = 232$, $0.002 < \pi_i < \Pr(i \in s) < 0.15$. For $n = 1,145$, $0.01 < \pi_i < 0.73$. For $n = 2,429$, $0.03 < \pi_i < 0.99$ with mean $\bar{\pi} = 0.26$ and standard deviation $Std(\pi_i) = 0.29$. In the latter case some of the units were drawn almost with certainty).

Some of the predictors considered for this study (see below) require the specification of either the sample model or the sample-complement model. We assumed for both models the third order polynomial regression,

$$y_k = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \beta_3 x_k^3 + \varepsilon_k \quad (7.1)$$

with independent residuals and constant variance. This model was found by Pfeffermann and Sverchkov (1999) to give a good fit to the ‘population’ (original sample) data with $R^2 = 0.61$ (see Figure 1), and it was found also to fit fairly well the sample data (with different coefficients) for several samples selected from this ‘population’. Notice, on the other hand, that with this strongly informative sampling scheme, it is unlikely that the sample model, the population model and the sample-complement model are all from the same family even if with different parameters. The present study enables therefore studying the performance of the various predictors when some or all of the three models are misspecified. This important robustness question is further examined by fitting simple regression models instead of the third order polynomial regressions that is, by omitting the second and third powers of the auxiliary variable. The only exception is the model dependent predictor \hat{Y}_1 (Equation 4.4) where no coherent estimator for the expectation $E_s(w_j | x_j)$ could be found when restricting to simple regression. (The method considered in Pfeffermann and Sverchkov (1999) for the estimation of this expectation assumes normality of the population model residuals. This is a valid assumption when fitting the third order polynomial regression model but is clearly violated when dropping the second and third powers of the auxiliary variable).

U.S. National Maternal and Infant Health Survey, 1988.

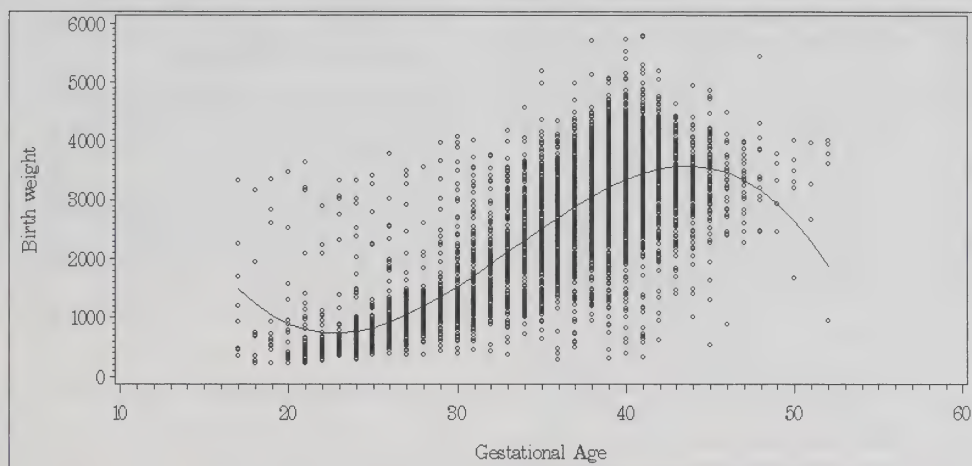
Model Fitted: $y_i = 17886 - 1827.7x_i + 61.2x_i^2 - 0.61x_i^3 + \varepsilon_i$ $\text{Var}(\varepsilon_i) = 603.2, R^2 = 0.61$

Figure 1. Scatterplot of Birth Weight against Gestational Age in 'Population' (original Sample), and Predicted Values Under 3rd Order Polynomial Regression.

The predictors considered for this study divide therefore into three groups. The first group consists of predictors that only use the sample y -values and the sampling weights. Included in this group are the Horvitz-Thompson estimator $\hat{Y}_{H-T} = \sum_{i \in s} w_i y_i$, the predictor \hat{Y}_{EI} defined by (5.2) and Hajek's estimator \hat{Y}_{Hajek} defined by (5.3). The second group consists of predictors that use the working model defined by (7.1). Included in this group are the two regression predictors \hat{Y}_1 and $\hat{Y}_{2, \text{Reg}}$ defined by (4.4) and (4.8) respectively, the bias corrected predictor \hat{Y}_3 defined by (4.12) and the GREG estimator defined by (4.14). The third group contains the same predictors as the second group (except for \hat{Y}_1 , see above), but based on the simple regression model (only the first power of x).

The MSEs of all the predictors considered in this study have been estimated by use of the two-step procedure described in section 6. However, because of computing time limitations, the MSE estimators were only computed for a random selection of 200 out of the 1,000 samples and are based on only 200 bootstrap samples from each pseudo population. For assessing the performance of the MSE estimators we computed the corresponding empirical MSEs based on the 1,000 samples selected from the study population. Thus, the 'true' MSE of a generic predictor \hat{Y} was computed as,

$$\text{MSE}(\hat{Y}) = \frac{1}{1,000} \sum_{r=1}^{1,000} (\hat{Y}_{(r)} - Y)^2 \quad (7.2)$$

where $\hat{Y}_{(r)}$ denotes the predictor computed from the r^{th} sample. Notice that since the population values are fixed, the MSE in (7.2) is the randomization MSE over all possible sample selections, which is what the estimator (6.2) is intended to estimate.

7.2 Results of Empirical Study

The main results of this study are exhibited in Tables 1.1 – 1.3 (one table for each sample size). The third column of each table shows for every predictor \hat{Y} the empirical bias, $[(\sum_{r=1}^R \hat{Y}_{(r)} / R) - Y]$, and the standard deviation (*Std*) of the empirical bias, computed as $[\sum_{r=1}^R (\hat{Y}_{(r)} - \bar{Y}_R)^2 / R^2]^{1/2}$; $\bar{Y}_R = \sum_{r=1}^R \hat{Y}_{(r)} / R$, $R = 1,000$. The next two columns show respectively the 'true' (empirical) RMSE (square root of Equation 7.2), and the square root of the mean of the corresponding Bootstrap estimators defined by (6.2).

The main conclusions from Tables 1.1 – 1.3 are as follows:

- 1- All the predictors considered for this study are virtually design unbiased with all three sample sizes, irrespective of the underlying working model. The predictor \hat{Y}_1 has a statistically significant bias when tested by use of the conventional t -statistic but the actual bias is negligible when compared to the true population total. (The predictor \hat{Y}_1 is the only predictor considered in this study that is not design consistent).

The next three comments refer to the RMSE of the various predictors.

- 2- The predictors in Groups 2 and 3 that use the auxiliary values perform much better than the predictors in Group 1, particularly for the smaller sample sizes. The predictors in Group 2 that employ the 3rd order polynomial regression model (7.1) perform better than the corresponding predictors in Group 3 that employ the simple regression model as the working model, but the differences diminish as the sample size increases.
- 3- An important result emerging from this study is that the predictors $\hat{Y}_{2, \text{Reg}}$ and \hat{Y}_{EI} (and also \hat{Y}_3 for the larger sample sizes), that only predict the y -values for units outside the sample indeed perform better than the other predictors in their respective groups (see also below). As surmised in Remark 7, this holds particularly with the larger sample sizes. Notice that the differences between $\hat{Y}_{2, \text{Reg}}$ and the GREG estimator for $n=1,145$ and $n=2,250$ are smaller under the polynomial model (Group 2) than under the simple regression model (Group 3), which is explained by the tight relationship between the study variable and auxiliary variables under the polynomial model. The predictor \hat{Y}_3 is less stable than $\hat{Y}_{2, \text{Reg}}$ for $n=232$ but for the other two sample sizes the two predictors perform similarly.
- 4- The predictor $\hat{Y}_{2, \text{Reg}}$ performs somewhat better than the model dependent predictor \hat{Y}_1 that employs the expectations $E(w_i | x_i)$ to adjust the sampling weights. We have no clear explanation for this result because as illustrated in Pfeffermann and Sverchkov (1999) using

the same data, adjusting the sampling weights improves the estimation of the regression coefficients very significantly.

Next consider the MSE estimators.

- 5- The MSE estimators developed in section 6 perform very well for all the predictors and with all the sample sizes. For the sample size $n=232$ there is a systematic under-estimation of the RMSE by up to 3%, which is explained by the fact that the pseudo population is in this case less variable than the actual study population (see Remark 9). The MSE estimators are almost unbiased for the other sample sizes with the largest difference between the estimated and true RMSE being again in the magnitude of 3%.

Another way of assessing the bias of the various predictors and their MSE estimation is by studying the coverage properties of confidence intervals defined by these predictors. Tables 2.1–2.3 compare the empirical percentage coverage of the standard confidence intervals $\hat{Y} \pm Z_{1-\alpha/2} \sqrt{\hat{MSE}}$ with the corresponding nominal percentages for selected values of α (one table for each sample size). The empirical percentages are somewhat erratic with $n=232$ sample units but they stabilize as the sample size increases, particularly with the use of the predictors in the second and third group. The empirical percentages are close to the nominal percentages with all the predictors when $n=2,250$.

Table 1.1
Bias, RMSE and Square Root of Mean of MSE Estimators, $n = 232$

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\hat{MSE}}$
1 No x -values	\hat{Y}_{H-T}	-4.5 (11.6)	365.1	355.0
	\hat{Y}_{EI}	1.5 (2.9)	91.1	89.8
	\hat{Y}_{Hajek}	1.7 (2.9)	93.0	91.6
2	\hat{Y}_1	4.4 (2.0)	64.0	63.0
3 rd order	$\hat{Y}_{2, \text{Reg}}$	3.5 (2.0)	63.4	62.4
polynomial	\hat{Y}_3	-0.3 (2.1)	65.4	65.0
regression	\hat{Y}_{GREG}	3.4 (2.1)	63.6	62.6
3 Simple Regression	$\hat{Y}_{2, \text{Reg}}$	-2.3 (2.2)	68.0	66.2
	\hat{Y}_3	-0.3 (2.2)	68.6	67.4
	\hat{Y}_{GREG}	-2.3 (2.2)	68.3	66.5

True 'population' total= 2710.7

Table 1.2
Bias, RMSE and Square Root of Mean of MSE Estimators, $n = 1,145$

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\text{MSE}}$
1 No x -values	\hat{Y}_{H-T}	-9.1 (5.0)	157.1	156.1
	\hat{Y}_{EI}	0.0 (1.1)	35.2	34.9
	\hat{Y}_{Hajek}	-0.1 (1.3)	39.5	39.3
	\hat{Y}_1	3.0 (0.9)	27.6	28.1
2 3 rd order polynomial regression	$\hat{Y}_{2, \text{Reg}}$	2.0 (0.9)	27.4	27.3
	\hat{Y}_3	0.5 (0.9)	27.4	27.7
	\hat{Y}_{GREG}	1.7 (0.9)	27.8	27.8
	$\hat{Y}_{2, \text{Reg}}$	0.0 (1.0)	28.3	28.7
3 Simple Regression	\hat{Y}_3	0.1 (1.0)	28.2	28.9
	\hat{Y}_{GREG}	0.0 (2.0)	29.1	29.6

True 'population' total= 2710.7

Table 1.3
Bias, RMSE and Square Root of Mean of MSE Estimators, $n=2,250$

Group	Predictor	Bias (Std)	RMSE	$\sqrt{\text{MSE}}$
1 No x -values	\hat{Y}_{H-T}	1.3 (2.7)	82.7	80.4
	\hat{Y}_{EI}	-0.2 (0.6)	18.5	18.8
	\hat{Y}_{Hajek}	0.1 (0.7)	23.5	23.8
	\hat{Y}_1	1.3 (0.5)	17.5	17.3
2 3 rd order polynomial regression	$\hat{Y}_{2, \text{Reg}}$	0.6 (0.5)	16.9	16.3
	\hat{Y}_3	-0.3 (0.5)	17.1	16.5
	\hat{Y}_{GREG}	0.5 (0.5)	17.9	18.3
	$\hat{Y}_{2, \text{Reg}}$	-0.3 (0.5)	17.3	16.8
3 Simple Regression	\hat{Y}_3	-0.3 (0.5)	17.7	17.3
	\hat{Y}_{GREG}	-0.2 (0.6)	18.8	18.3

True 'population' total= 2710.7

Table 2.1
Nominal and Empirical Percentage Coverage of Confidence Intervals, $n = 232$

Group	Predictor	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
1 No x -values	\hat{Y}_{H-T}	2.5	3.5	5.5	10.0	90.0	97.0	99.0	99.5
	\hat{Y}_{EI}	0.5	2.0	4.0	8.0	88.5	91.5	95.5	98.0
	\hat{Y}_{Hajek}	0.5	2.0	4.0	8.0	88.5	91.5	95.5	98.0
	\hat{Y}_1	0.0	0.0	1.5	6.5	86.0	90.5	92.5	97.5
2 3 rd order polynomial regression	$\hat{Y}_{2, \text{Reg}}$	0.0	0.0	2.0	7.0	85.0	90.5	93.5	98.0
	\hat{Y}_3	0.0	0.5	2.5	6.5	87.5	91.0	95.0	98.5
	\hat{Y}_{GREG}	0.0	0.0	2.0	7.0	85.0	90.5	93.5	98.0
	$\hat{Y}_{2, \text{Reg}}$	0.0	1.0	2.5	7.0	87.0	91.5	97.5	98.0
3 Simple Regression	\hat{Y}_3	0.0	1.0	2.5	7.0	86.0	91.5	96.5	98.0
	\hat{Y}_{GREG}	0.0	1.0	2.5	7.0	86.5	91.5	97.0	98.0

Table 2.2
Nominal and Empirical Percentage Coverage of Confidence Intervals, $n = 1,145$

Group	Predictor	1	2.5	5.0	10.0	90.0	95.0	97.5	99.0
1 No x -values	\hat{Y}_{H-T}	4.0	7.0	9.0	13.5	95.5	98.0	98.5	99.5
	\hat{Y}_{EI}	3.0	5.0	8.0	12.5	92.5	95.5	99.5	100.0
	\hat{Y}_{Hajek}	3.5	5.0	9.5	12.5	92.5	96.0	99.5	100.0
	\hat{Y}_1	0.5	2.0	5.0	7.5	86.5	93.5	96.0	97.0
2 3 rd order polynomial regression	$\hat{Y}_{2,Reg}$	0.5	3.0	6.0	9.0	86.5	94.5	96.5	97.0
	\hat{Y}_3	0.5	2.0	6.0	9.5	88.0	94.0	97.0	98.0
	\hat{Y}_{GREG}	0.5	3.0	5.0	9.0	86.5	94.0	96.5	98.0
	$\hat{Y}_{2,Reg}$	0.5	3.0	6.0	11.0	90.0	93.0	97.0	99.5
3 Simple Regression	\hat{Y}_3	0.5	2.5	5.5	10.5	90.0	94.0	97.0	99.5
	\hat{Y}_{GREG}	1.0	3.0	6.0	11.0	90.5	94.0	97.5	99.0

Table 2.3
Nominal and Empirical Percentage Coverage of Confidence Intervals, $n = 2,250$

Group	Predictor	1.0	2.5	5.0	10.0	90.0	95.0	97.5	99.0
1 No x -values	\hat{Y}_{H-T}	0.5	1.0	5.5	11.0	95.0	97.5	99.0	99.5
	\hat{Y}_{EI}	1.0	3.0	5.5	9.0	91.5	96.0	99.0	99.5
	\hat{Y}_{Hajek}	1.0	2.5	5.5	9.0	93.0	97.0	98.5	99.5
	\hat{Y}_1	0.5	2.0	5.0	9.0	91.0	94.5	96.5	97.5
2 3 rd order polynomial regression	$\hat{Y}_{2,Reg}$	0.5	2.5	6.5	10.5	90.5	94.5	96.5	98.0
	\hat{Y}_3	0.5	2.0	7.5	12.5	91.5	95.5	96.5	97.5
	\hat{Y}_{GREG}	0.5	2.0	6.0	11.0	91.0	94.5	96.0	98.0
	$\hat{Y}_{2,Reg}$	1.0	3.0	6.0	11.0	91.0	95.0	97.5	99.0
3 Simple Regression	\hat{Y}_3	1.0	2.0	6.0	12.0	90.0	95.0	97.5	98.0
	\hat{Y}_{GREG}	0.0	1.5	5.0	11.5	91.5	95.0	97.5	99.0

As implied by the theoretical developments of this article and illustrated in the empirical study, predicting only the y -values for units outside the sample employing the sample-complement model yields better predictors for the population total than predicting all the population values by use of the population model, as implicitly implemented when using the GREG or Hajek's estimators. Clearly, the differences are only appreciable when the sampling fractions are not negligible.

In order to highlight this point further, we present in Table 3 the mean prediction error (mpe) in the original scale (grams) over the 1,000 samples when predicting the sample-complement values;

$$mpe = \sum_{r=1}^{1,000} \left[\sum_{j \in S_r} (\hat{y}_j - y_j) / (N - n) \right] / 1,000$$

where S_r defines the r^{th} selected sample. The mpe's are shown for three predictors, all utilizing the working model (7.1) and thus having the general form, $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j + \hat{\beta}_2 x_j^2 + \hat{\beta}_3 x_j^3$, $j \in s$. For the first predictor the vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ is estimated by OLS, which corresponds to the use of the sample model; for the second predictor β is estimated by the probability weighted estimator \hat{B}_{pw} , that corresponds to the use of the population model whereas for the third predictor β is estimated by the estimator \hat{B}_c which is computed similarly to \hat{B}_{pw} but with weights $(w_i - 1)$, that corresponds to the use of the sample-complement model.

Table 3
Mean Prediction Errors and Std of Means (in brackets) Under Three Prediction Models

Sample size	Sample Model	Population model	Sample-Complement model
232	329.0 (2.2)	10.3 (2.3)	4.3 (2.3)
1,145	375.0 (0.9)	37.7 (1.1)	2.4 (1.1)
2,250	387.5 (0.6)	85.8 (0.7)	0.9 (0.8)

The clear conclusion emerging from Table 3 is that the use of either the population model or the model holding for units in the sample for the prediction of y -values of units outside the sample can result in appreciable biases. Notice that the bias induced by use of the population model increases as the sampling fraction increases, which agrees with the previous discussion asserting that the difference between the sample and sample-complement models only shows up with relatively large sample sizes (see Comment 2).

8. CONCLUDING REMARKS

In this article we use the sample and sample-complement distributions for developing *design consistent* predictors of finite population totals. Known predictors in common use are shown to be special cases of the present theory. The MSEs of the new predictors are estimated by a combination of an inverse sampling algorithm and a resampling method. As supported by theory and illustrated in the empirical study, predictors of finite population totals that only require the prediction of the outcome values for units outside the sample perform better than predictors in common use even under a design based framework, unless the sampling fractions are very small. The MSE estimators are shown to perform well both in terms of bias and when used for the computation of confidence intervals for the population totals. Further experimentation with this kind of predictors and MSE estimation is therefore highly recommended.

ACKNOWLEDGEMENT

The authors would like to thank the associate editor and two referees for very constructive comments.

REFERENCES

- BREWER, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumptions of an underlying stochastic process. *Australian Journal of Statistics*. 5, 93-105.
- BREWER, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*. 25, 205-212.
- CHAMBERS, R.L., DORFMAN, A. and SVERCHKOV, M. (2003). Nonparametric regression with complex survey data. In, *Analysis of Survey Data*, (Eds. C. Skinner and R. Chambers). New York: John Wiley & Sons, Inc. 151-174.
- FULLER, W. (2003). Statistical analysis from complex survey data. Tutorial presented at the International Statistical Institute meeting, Berlin, Germany. Slides of the Tutorial appear in <http://cssm.iastate.edu/academic/staff/fuller.html>.
- HANSEN, M.H., MADOW, W.G. and TEPPIG, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association*. 78, 776-807.
- KIM, D.H. (2002). Bayesian and empirical Bayesian analysis under informative sampling. *Sankhyā* B. 64, 267-288.
- KORN, E.L., and GRAUBARD, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*. 49, 291-295.
- PATAK, Z., HIDIROGLOU, M. and LAVALLÉE, P. (2000). The methodology of the Workplace and Employee Survey. *Proceedings of the Second International Conference on Establishment Surveys*, June 17-21, 2000, Buffalo, New York, American Statistical Association. 223-232.
- PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*. 61, 317-337.
- PFEFFERMANN, D., and KRIEGER, A.M. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*. 13, 123-142.
- PFEFFERMANN, D., KRIEGER, A.M. and RINOTT, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*. 8, 1087-1114.
- PFEFFERMANN, D., and SVERCHKOV, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, Series B. 61. 166-186.
- PFEFFERMANN, D., and SVERCHKOV, M. (2003a). Fitting generalized linear models under informative probability sampling. In *Analysis of survey Data*, (Eds. C. Skinner and R. Chambers). New York: John Wiley & Sons, Inc. 175-195.
- PFEFFERMANN, D., and SVERCHKOV, M. (2003b). Small area estimation under informative sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association (to appear).

- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*. 57, 377-387.
- SÄRNDAL, C.E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*. 67, 639-650.

Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs

LEONARDO GRILLI and MONICA PRATESI¹

ABSTRACT

Multilevel models are often fitted to survey data gathered with a complex multistage sampling design. However, if such a design is informative, in the sense that the inclusion probabilities depend on the response variable even after conditioning on the covariates, then standard maximum likelihood estimators are biased. In this paper, following the Pseudo Maximum Likelihood (PML) approach of Skinner (1989), we propose a probability-weighted estimation procedure for multilevel ordinal and binary models which eliminates the bias generated by the informativeness of the design. The reciprocals of the inclusion probabilities at each sampling stage are used to weight the log-likelihood function and the weighted estimators obtained in this way are tested by means of a simulation study for the simple case of a binary random intercept model with and without covariates. The variance estimators are obtained by a bootstrap procedure. The maximization of the weighted log-likelihood of the model is done by the NLMIXED procedure of the SAS, which is based on adaptive Gaussian quadrature. Also the bootstrap estimation of variances is implemented in the SAS environment.

KEY WORDS: Informative design; Multilevel ordinal model; Multistage sampling; Pseudo Maximum Likelihood; Weighting.

1. INTRODUCTION

Multilevel models for ordinal responses, including binary responses as a special case, are frequently used in many areas of research for modelling hierarchically clustered populations. In fact, both in human and biological sciences, the status or the response of a subject may often be classified in two categories or in a set of ordered categories (ordinal or graded scale). At the same time, subjects are observed clustered in groups (*e.g.*, schools, firms, clinics, geographical areas). The hierarchical population structure is often also employed to design multistage sampling schemes, with unequal selection probabilities at some or all the stages of the sampling process. In the multilevel analysis of survey data, complex sampling schemes are often ignored even if they may cause the violation of the basic assumptions underlying multilevel models. In fact, in complex sampling designs both the subjects and the clusters at all levels could be selected with probabilities that, even conditionally on the covariates, do depend on the response variable; in other words, the sampling design might be informative.

For data that are clustered and obtained by multistage informative designs, proposals for fitting multilevel models have been formulated mainly for the case of continuous response variables. In particular, Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998) propose probability-weighting procedures of first and second level units that adjust for the effect of an informative design on the

estimation in two-level models with a continuous response variable. The method, known as Pseudo Maximum Likelihood (PML), consists in writing down a closed form expression for the census likelihood, estimating the log-likelihood function and then maximizing the estimated function numerically. The method needs the sampling weights for the sampled elements and clusters at all levels. The authors also develop appropriate 'sandwich' estimators for the variances of the estimators.

The work of Pfeffermann *et al.* (1998) is mainly concerned with the implementation of the PML principle in the IGLS (Iterative Generalised Least Squares) algorithm (Goldstein 1986), which is suitable for linear multilevel models. The probability-weighted IGLS algorithm is available in the widespread package MLwiN (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse and Draper 1999). However, the extension to nonlinear models is not trivial. For the nonlinear case the developers of MLwiN implemented a weighting procedure that parallels the one used for linear models with some ad hoc solution for the level 1 variation: for example, for binary responses the subject-level weights are included in the binomial denominator. The proposed method is straightforward to implement, but its properties have not been investigated yet. Moreover Renard and Molenberghs (2002) report the case of an application where the aforementioned algorithm for weighting in multilevel binary models did not converge or yielded implausible results.

¹ Leonardo Grilli, Dipartimento di Statistica, Università di Firenze. E-mail: grilli@ds.unifi.it; Monica Pratesi, Dipartimento di Statistica e Matematica applicata all'Economia, Università di Pisa. E-mail: m.pratesi@ec.unipi.it.

The simulation study which we will use to judge the performance of the PML estimators will closely follow the lines of Pfeiffermann *et al.* (1998), since they use a similar approach for the linear model, so that some interesting comparisons are possible. However, when making the comparisons it should always be kept in mind that, while in the two-level linear model the two variance components can be estimated separately, in the two-level binary model only a ratio of the two variance components is estimable, as discussed further on.

A recent paper which deals with the estimation of variance components is Korn and Graubard (2003), whose work is motivated by the substantial bias showed in small samples by several weighted estimators of variance components proposed to adjust for informative designs (Graubard and Korn 1996). Though the topic is same, the work of Korn and Graubard is different from ours in many respects: a) As Pfeiffermann *et al.* (1998), they consider only the linear multilevel model. b) In the context of the linear multilevel model, they focus on unbiased estimation of the variance components in small samples: in fact they propose some estimators for the variance components and only sketch how to derive similar estimators for the linear model with covariates, but without testing their performance. Anyway, the extension to nonlinear multilevel models is not trivial. c) The main estimators proposed by Korn and Graubard (2003), which are in closed form, showed good performance even in small samples. However they rely on the pairwise joint inclusion probabilities. When such probabilities are not available, which is often the case in practice, the authors propose a variant whose bias is substantial when the number of sampled clusters is moderate (33 in their simulation plan). In contrast, the PML method adopted in our work do not require joint inclusion probabilities. d) The informative design used by Korn and Graubard (2003) for their simulation study is quite different from ours: in fact, in their design the undersampling of the units depends on whether the model's random errors are greater than a certain threshold in absolute value, while in our design the criterion depends on whether the random errors are high or low. Therefore a comparison of the results is difficult.

The wide use of nonlinear multilevel models in many fields of application urges for a general and reliable weighted estimation method, which should be both effective and simple to implement, preferably in the framework of a standard statistical software. The present paper represents a contribution in this direction.

It is worth to note that the PML method we exploit is quite general, so it can be applied to a wide range of models. In the paper the focus is on models for ordinal and binary responses, since they are very common and can be represented as a linear model for the latent response

endowed with a set of thresholds (see section 2), facilitating the comparison with the existing results for the linear model. However the description of the PML approach is absolutely general and the estimation technique based on the NLMIXED procedure of SAS (reported in Appendix A) is easy to generalize.

The structure of the paper is as follows. Basic definitions for the multilevel ordinal model are set out in section 2, while in section 3 the general PML approach is described, along with some details for fitting the model using SAS NLMIXED. In section 4 the properties of the various estimators for the random intercept binary model are evaluated by a simulation study. Section 5 concludes with some final remarks.

2. THE MULTILEVEL ORDINAL MODEL

In order to ease the comparison with the results concerning the linear model (Pfeiffermann *et al.* 1998; Korn and Graubard 2003), it is useful to write the ordinal model in terms of a latent linear model endowed with a set of thresholds. Suppose that an observed ordinal response variable Y , with $k = 1, 2, \dots, K$ levels, is generated, through a set of thresholds, by a latent continuous variable \tilde{Y} following a variance component model (Hedeker and Gibbons 1994):

$$\tilde{Y}_{ij} = \beta' \mathbf{x}_{ij} + \omega u_j + \varepsilon_{ij}, \quad (1)$$

with $i = 1, 2, \dots, N_j$ elementary units (subjects) for the j -th cluster ($j = 1, 2, \dots, M$). In (1) \mathbf{x}_{ij} is a covariate vector and β is the corresponding vector of slopes; the random variables ε_{ij} and u_j are the disturbances, respectively at the first (subject) and second (cluster) level; and ω^2 is the second level variance component.

For the disturbances of model (1) we make the standard assumptions, *i.e.*, a) the ε_{ij} 's are iid with zero mean and unknown variance σ^2 ; b) the u_j 's are Gaussian iid with zero mean and unit variance; c) the ε_{ij} 's and u_j 's are mutually independent.

Note that model (1) leads to the simplest case of a multilevel ordinal model, with just two levels and a single random effect on the intercept; the extension to three or more levels and to multiple random effects is straightforward in principle (Gibbons and Hedeker 1997), but the complications in the formulae suggest to consider only the simplest case, which is sufficient for the discussion of the main conceptual issues.

The observed ordinal variable Y is linked to the latent one \tilde{Y} through the following relationship:

$$\{Y_{ij} = k\} \Leftrightarrow \{\gamma_{k-1} < \tilde{Y}_{ij} \leq \gamma_k\},$$

where the thresholds satisfy $-\infty = \gamma_0 \leq \gamma_1 \leq \dots \leq \gamma_{K-1} \leq \gamma_K = +\infty$. Therefore, conditional on u_j , the model probability for subject i of cluster j is

$$\begin{aligned} P(Y_{ij} = k | u_j) &= P(\gamma_{k-1} < \tilde{Y}_{ij} \leq \gamma_k | u_j) \\ &= P(\tilde{Y}_{ij} \leq \gamma_k | u_j) - P(\tilde{Y}_{ij} \leq \gamma_{k-1} | u_j), \end{aligned} \quad (2)$$

with

$$\begin{aligned} P(\tilde{Y}_{ij} \leq \gamma_k | u_j) &= P(\varepsilon_{ij} \leq \gamma_k - [\beta' \mathbf{x}_{ij} + \omega u_j] | u_j) \\ &= F\left(\frac{\gamma_k}{\sigma} - \left[\frac{1}{\sigma} \beta' \mathbf{x}_{ij} + \frac{\omega}{\sigma} u_j\right]\right) \\ &= F(\gamma_{\sigma,k} - [\beta'_{\sigma} \mathbf{x}_{ij} + \omega_{\sigma} u_j]), \end{aligned} \quad (3)$$

where $F(\cdot)$ is the distribution function of the standardized first level error term ε_{ij}/σ . All the model parameters are defined in terms of the unknown σ , the standard deviation of the first level error term, so only the ratios of the model parameters to the standard deviation of the first level error term are identifiable; we use the notation ψ_{σ} to indicate that the latent model parameter ψ is in σ units, *i.e.*, $\psi_{\sigma} \equiv \psi/\sigma$. Note that $F(\cdot)$ is also the inverse of the link function of the ordinal model: for example, the standard Gaussian distribution function yields the ordinal probit model.

As for identification, if β_{σ} includes the intercept, the estimable thresholds are $K-2$; so it is customary to set $\gamma_{\sigma,1} = 0$. Alternatively, if the intercept is fixed to zero all the $K-1$ thresholds are estimable.

Now let θ denote the vector of all estimable parameters, which include β_{σ} , ω_{σ} and $K-2$ thresholds $\{\gamma_{\sigma,k}; k=2, \dots, K-1\}$ ($\gamma_{\sigma,1}$ is fixed to zero to insure identifiability). The conditional likelihood for subject i of cluster j is

$$L_{ij}(\theta | u_j) = \prod_{k=1}^K [P(Y_{ij} = k | u_j)]^{d_{ijk}}, \quad (4)$$

where $P(Y_{ij} = k | u_j)$ is defined by (2) and (3), while d_{ijk} is the indicator function of the event $\{Y_{ij} = k\}$. Then the marginal likelihood for cluster j is

$$L_j(\theta) = \int_{-\infty}^{+\infty} \prod_{i=1}^{N_j} L_{ij}(\theta | u) \varphi(u) du,$$

where φ is the standard Gaussian density function. Finally, the overall marginal likelihood is

$$L(\theta) = \prod_{j=1}^M L_j(\theta). \quad (5)$$

3. PROBABILITY-WEIGHTED ESTIMATION

3.1 Pseudo Maximum Likelihood (PML) Estimators

Suppose that the whole population of M clusters (level 2 units) with N_j elementary units (subjects or level 1 units) per cluster is not observed; instead the following two-stage sampling scheme is used:

- first stage: m clusters are selected with inclusion probabilities π_j ($j = 1, \dots, M$);
- second stage: n_j elementary units are selected within the j -th selected cluster with probabilities π_{ij} ($i = 1, \dots, N_j$).

The unconditional sample inclusion probabilities are then $\pi_{ij} = \pi_{ij} \pi_j$.

When the sampling mechanism is informative, *i.e.*, the π_j and/or the π_{ij} depend on the model disturbances and hence on the response variable, the maximum likelihood estimator of the parameters of the multilevel ordinal model defined in section 2 may be seriously biased.

A standard solution to this problem is provided by the Pseudo Maximum Likelihood (PML) approach (Skinner 1989). However in the context of multilevel models the implementation of the PML approach is complicated by the fact that the population log-likelihood is not a simple sum of elementary unit contributions, but rather a function of sums across level 2 and level 1 units. This can be seen by writing the logarithm of the likelihood (5) as follows:

$$\log L(\theta) = \sum_{j=1}^M \log \int_{-\infty}^{+\infty} \left[\exp \left\{ \sum_{i=1}^{N_j} \log L_{ij}(\theta | u) \right\} \right] \varphi(u) du. \quad (6)$$

A design consistent estimate of the population log-likelihood (6) can be obtained applying the Horvitz-Thompson principle, *i.e.*, replacing each sum over the level 2 population units j by a sample sum weighted by $w_j \equiv 1/\pi_j$ and each sum over the level 1 units i by a sample sum weighted by $w_{ij} \equiv 1/\pi_{ij}$:

$$\log \hat{L}(\theta) =$$

$$\sum_j w_j \log \int_{-\infty}^{+\infty} \left[\exp \left\{ \sum_i w_{ij} \log L_{ij}(\theta | u) \right\} \right] \varphi(u) du, \quad (7)$$

where \sum^s denotes a sum over sample units.

Note that inserting the weights in the log-likelihood implies the use of a design consistent estimator of the population score function. In fact, the population score function $U(\theta) \equiv \partial/\partial\theta \log L(\theta)$ can be written as

$$\sum_{j=1}^M \frac{\int_{-\infty}^{+\infty} \left[\exp \left\{ \sum_{i=1}^{N_j} \log L_{ij} \right\} \right] \cdot \left\{ \sum_{i=1}^{N_j} \frac{\partial}{\partial \theta} \log L_{ij} \right\} \varphi(u) du}{\int_{-\infty}^{+\infty} \exp \left\{ \sum_{i=1}^{N_j} \log L_{ij} \right\} \varphi(u) du}, \quad (8)$$

where $L_{ij} = L_{ij}(\theta|u)$, whose corresponding Horvitz-Thompson estimator $\hat{U}(\theta)$ is

$$\sum_j \frac{\int_{-\infty}^{+\infty} \left[\exp \left\{ \sum_i w_{ij} \log L_{ij} \right\} \right] \cdot \left\{ \sum_i w_{ij} \frac{\partial}{\partial \theta} \log L_{ij} \right\} \varphi(u) du}{\int_{-\infty}^{+\infty} \exp \left\{ \sum_i w_{ij} \log L_{ij} \right\} \varphi(u) du}, \quad (9)$$

which equals the score obtained by differentiating the probability-weighted loglikelihood (7).

Under mild conditions, the solution $\hat{\theta}_{\text{PML}}$ to the estimating equations $\hat{U}(\theta) = \mathbf{0}$ is design consistent for the finite population maximum likelihood estimator $\hat{\theta}$ which, in turn, is model-consistent for the super-population parameter θ : therefore $\hat{\theta}_{\text{PML}}$ is a consistent estimator of θ with respect to the mixed design-model distribution (Pfeffermann 1993).

Note that general probability-weighted estimators for nonlinear multilevel models can also be devised by weighting suitable estimating functions, as in the work of Singh, Folsom and Vaish (2002) in the context of small area estimation.

The implementation of the PML approach requires the knowledge of the inclusion probabilities at both levels. Using only second level weights or only first level weights may be insufficient or may even worsen the situation, as shown by our simulations.

3.2 Scaling the Weights

A controversial issue discussed in Pfeffermann *et al.* (1998) and Korn and Graubard (2003) is the scaling of the weights to obtain estimators with little bias even in small samples. Obviously, scaling is not relevant for the level 2 weights, since from (7) and (9) it is clear that multiplying the w_j 's by a constant does not change the PML estimates (it simply inflates the information matrix by that constant). On the contrary, scaling the level 1 weights may have important effects on the small sample behavior of the PML estimator. In the simulation study discussed in section 4 we present the results for the following type of scaling (named 'scaling method 2' in Pfeffermann *et al.* 1998):

$$w_{ij}^{\text{scaled}} = \frac{w_{ij}}{\bar{w}_j}, \quad (10)$$

where $\bar{w}_j = (\sum_i w_{ij})/n_j$, so that, for the j -th cluster, the sum of the scaled weights equals the cluster sample size n_j . In the present paper we do not wish to discuss the relative merits of the various scaling methods, so we limit our simulations to scaled weights (10), which have an intuitive meaning and showed good performance in the study of Pfeffermann *et al.* (1998), although they may yield a substantial bias with certain designs, as discussed in Korn and Graubard (2003). The topic will be broached again in section 4.

3.3 Estimation Technique

The maximization of the weighted log-likelihood (7) involves the computation of several integrals which do not have a closed-form solution, so a numerical approximation technique is required. When the dimensionality of the integrals is low, a simple and very accurate technique is Gaussian quadrature, which is based on a summation over an appropriate set of points. The NLMIXED procedure of SAS (SAS Institute 1999) is a general procedure for fitting nonlinear random effects models using adaptive Gaussian quadrature. Various optimization techniques are available to carry out the maximization; the default, used in the simulations of section 4, is a dual quasi-Newton algorithm, where dual means that the upgrading concerns the Cholesky factor of an approximate Hessian (SAS Institute 1999).

Though the NLMIXED procedure does not include an option for PML estimation, it is still possible to insert the weights in the likelihood, using different tricks for level 1 and level 2 weights, as explained in Appendix A.

3.4 Variance Estimation

In standard maximum likelihood the estimation of the covariance matrix of the estimators is obtained by inverting the information matrix. However this conventional estimator is not appropriate when using the PML method since it does not take into account the variability stemming from the sampling design. To get a more reliable covariance matrix Skinner (1989) proposed the use of a robust 'sandwich' estimator, which is employed also by Pfeffermann *et al.* (1998).

As noted in section 3.3, the NLMIXED procedure of SAS allows to fit the model with the PML approach, but the estimated covariance matrix, which is obtained by inverting the information matrix, is likely to be misleading in order to appreciate the actual variability of PML estimators. In the SAS framework the derivation of 'sandwich' estimators is not trivial. However, a simple and effective solution, requiring a bit of programming, is to empirically estimate the variance through the bootstrap technique for finite populations (Särndal, Swensson and Wretman 1992), which consists of the following steps: a) using the sample data, an artificial finite population is constructed, assumed to mimic

the real population; b) a series of independent bootstrap samples is drawn from the artificial finite population and for each bootstrap sample an estimate of the target parameter is calculated; c) the bootstrap variance estimate is obtained as the variance of the observed distribution of the bootstrap estimates.

The artificial finite population can be generated in the following way: i) for the j -th sampled cluster, each of the n_j sampled elementary units is replicated w_{ij} times, rounding the weight to the nearest integer, obtaining an artificial cluster of about N_j elementary units; ii) each of the m artificial clusters is replicated w_j times, rounding the weight to the nearest integer, obtaining an artificial population of about M clusters. Then the samples are selected from the artificial population in the following way: i) m clusters are resampled with probability proportional to π_j ; ii) for the j -th resampled cluster, n_j elementary units are resampled with probability proportional to π_{ij} .

When the sampling fraction m/M is low, most of the variance is due to the sampling of the clusters, so the bootstrap procedure described above could be simplified by omitting the steps concerning the elementary units, *i.e.*, step i) in the construction of the artificial population and step ii) in the resampling process.

A simpler resampling technique for variance estimation, considered by Korn and Graubard (2003), is the jackknife. In the case of clustered designs the technique entails the calculation of the variance from the set of point estimates obtained by deleting one cluster at a time, though the performance of the jackknife with correlated data is not always satisfactory (Shao and Tu 1995). In our simulation study the jackknife variance estimator seems unreliable, so it is not used. Further research is needed to fully evaluate the potentialities of the jackknife by testing some suitable modifications of the technique.

4. SIMULATION STUDY

4.1 Design of Experiment

The experiment reflects the two-stage scheme assumed for the observed variables: first, the finite population values are generated from the adequate superpopulation model (stage I) and then an informative or non-informative sample is selected from the finite population (stage II), with one sample per population. The two-stage selection scheme was repeated 1,000 times for each combination of sample size and type of informativeness. In order to compare our results with the ones obtained for the multilevel linear model, the experiment has been designed following the example of Pfeffermann *et al.* (1998, section 7).

The simulation study focussed on a simple instance of the model defined in section 2, namely the random intercept probit binary model, which has only two categories for the response variable (*i.e.*, $K=2$) and one cluster-level Gaussian random error. To parallel the study of Pfeffermann *et al.* (1998) the main simulation plan refers to the model without covariates, but some additional simulations are conducted to assess the performance of the estimators in the model with one cluster-level covariate and one subject-level covariate.

The values of the binary response variable Y_{ij} were generated using the following two-stage scheme which parallels the one of Pfeffermann *et al.* (1998):

- Stage I. Finite population values Y_{ij} ($j=1, \dots, M; i=1, \dots, N_j$) were obtained by first generating a value from the superpopulation latent model $\tilde{Y}_{ij} = \beta + u_j + \varepsilon_{ij}$, with $u_j \sim N(0, \omega^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$, and then putting $Y_{ij} = 0$ if $\tilde{Y}_{ij} \leq 0$ or $Y_{ij} = 1$ if $\tilde{Y}_{ij} > 0$ (recall that the binary model has only one threshold which is set to zero to guarantee identifiability). The latent model parameter values employed in the simulation are $\beta = 0$, $\omega^2 = 0.2$ and $\sigma^2 = 0.5$, so that the parameters estimable from the binary model are $\beta_\sigma = \beta/\sigma = 0$ and $\omega_\sigma = \omega/\sigma = 0.632$ (see expression (3)). The hierarchical structure of the population comprises $M = 300$ clusters, while the cluster sizes N_j were determined by $N_j = 75 \exp(\tilde{u}_j)$, with \tilde{u}_j generated from $N(0, \omega^2)$, truncated below by -1.5ω and above by 1.5ω . As a result, in our population N_j lies in the range $[38, 147]$ with mean around 80.
- Stage II. Once the finite population values were obtained, we adopted one of the following sampling schemes:
 - (a) *Informative at both levels*: first, m clusters were selected with probability proportional to a 'measure of size' X_j , *i.e.*, $\pi_j = mX_j / \sum_{j=1}^M X_j$; the measure X_j was determined in the same way as N_j but with \tilde{u}_j replaced by u_j , the random effect at level 2. The elementary units in the j -th sampled cluster were then partitioned into two strata according to whether $\varepsilon_{ij} > 0$ or $\varepsilon_{ij} \leq 0$ and simple random samples of sizes $0.25n_j$ and $0.75n_j$ were selected from the respective strata. The sizes n_j were either fixed, $n_j = n_0$, or proportional to N_j .
 - (b) *Informative only at level 2*: the scheme is the same as the previous one, except that simple random sampling was employed for the selection of level 1 units within each sampled cluster.

- (c) *Non-informative*: the scheme is the same as the previous one, except that the size measure X_j was set equal to N_j .

The simulation study included samples with $m = 35$ clusters and varying numbers of elementary units: large samples with fixed size $n_j = n_0 = 38$ and proportional allocation $n_j = 0.4N_j$, and small samples with fixed size $n_j = n_0 = 9$ and proportional allocation $n_j = 0.1N_j$ (mean of about 9).

The simulation study was carried out entirely within the SAS System (SAS Institute 1999), writing specific code with the macro language. The models were fitted with the NLMIXED procedure (see Appendix A), using 10-point adaptive Gaussian quadrature with a dual quasi-Newton algorithm, which reached convergence in a few iterations. As explained in Appendix A, to avoid gross rounding errors the level 2 weights were pre-multiplied by a factor $k = 10,000$ and the estimated covariance matrix was then multiplied by the same factor.

4.2 Results

The results of the simulations are shown in Tables 1 and 2. For each sampling design the behavior of the point estimators of the intercept β_0 and the second level standard deviation ω_σ is summarized by the mean and standard deviation of their Monte Carlo sampling distribution. The point estimators under study are the standard maximum likelihood unweighted estimator and the following three weighted versions of it: a) *cluster-level weighted*: the weights are only at level 2 (i.e., varying w_j 's and constant $w_{i|j}$'s); b) *unscaled fully weighted*: the weights are at both levels and the level 1 weights are unscaled; c) *scaled fully weighted*: the weights are at both levels and the level 1 weights are scaled according to (10), i.e., 'scaling method 2' of Pfeffermann *et al.* (1998).

Our results are shown and discussed according to the following three scenarios: 1) *Base scenario*: the sampling design is non-informative. In this situation all the basic assumptions underlying the random intercept binary model are fulfilled, so this case can be assumed as a benchmark for judging the subsequent results. 2) *Informative/Unweighted scenario*: the sampling design is informative, while the estimator is unweighted. In this situation the basic assumptions underlying the random intercept binary model are violated because of the informativeness of the design and no adjustment is used. 3) *Informative/Weighted scenario*: the sampling design is informative and the estimator is weighted. Also in this case the basic assumptions underlying the random intercept binary model are violated, but the weights are introduced as a tentative adjustment for the bias of the standard estimator.

4.2.1 Base Scenario

When the sampling design is non-informative the standard maximum likelihood unweighted estimator is asymptotically unbiased (Tables 1 and 2: rows 9-12, column 1). However for small samples ($n_j = 9$ and $n_j 0.1N_j$) there is an appreciable negative bias in the estimation of ω_σ .

If the weights are introduced when there is no need to adjust for the effect of the design (Tables 1 and 2: rows 9-12, columns 2-4), we face a slight increase in the variability of the estimators, which is more pronounced when the unscaled fully weighted estimator is used in small samples. Note that, still in small samples, the unscaled fully weighted estimator of ω_σ is upward biased.

4.2.2 Informative/Unweighted Scenario

The informativeness of the sampling design produces biased and unstable estimates. The bias is still evident for large samples (Tables 1 and 2: rows 1-8, column 1). The conclusions are the same for both types of informative designs, though the bias tends to have a different sign. Moreover the informativeness of the design inflates the variability of the standard estimator with respect to the base scenario: in particular, when the design is informative at both levels the standard error of the estimator of β_0 is doubled.

4.2.3 Informative/Weighted Scenario

Estimation of β_0

The results in Table 1 show that, when the design is informative, the weighted-based adjustment is effective in removing the bias in the estimation of β_0 .

Particularly, when the design is informative only at level 2 (Table 1: rows 5-8, columns 2-4) and the weights are introduced only at this level (cluster-level weighted estimator), the bias in the estimation is corrected with no important increase in the sampling variance. The result is valid also for fully weighted estimators (unscaled or scaled). The bias correction works for small samples too.

When the design is informative at both levels (Table 1: rows 1-4, columns 2-4) and the weights are introduced at both levels (fully weighted estimators), the bias in the estimation of β_0 is corrected. Moreover, the fully weighted estimators have smaller sampling variance than the unweighted counterpart, except for the unscaled version in small samples. The scaled version is preferable especially in small samples, since it allows to achieve an unbiased estimator with a substantial lower sampling variance. It should be noted that when the design is informative at both levels, the cluster-level weighted estimator is worse than the standard unweighted estimator.

Table 1

Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Intercept (true value 0, number of replicates 1,000)

Sampling design	Unweighted estimator	Weighted estimators		
		Cluster-level weighted	Unscaled fully weighted	Scaled fully weighted
Informative at both levels				
Fixed size $n_j = 38$	-0.120 (0.212)	-0.411 (0.202)	0.014 (0.193)	0.015 (0.188)
Prop. size $n_j = 0.4N_j$	-0.163 (0.212)	-0.453 (0.200)	0.018 (0.190)	0.021 (0.183)
Fixed size $n_j = 9$	-0.214 (0.204)	-0.512 (0.190)	-0.062 (0.258)	0.000 (0.185)
Prop. size $n_j = 0.1N_j$	-0.164 (0.220)	-0.450 (0.209)	-0.074 (0.294)	0.008 (0.203)
Informative only at cluster level (level 2)				
Fixed size $n_j = 38$	0.281 (0.169)	0.018 (0.168)	0.017 (0.170)	0.017 (0.169)
Prop. size $n_j = 0.4N_j$	0.274 (0.169)	0.014 (0.178)	0.014 (0.182)	0.014 (0.181)
Fixed size $n_j = 9$	0.274 (0.187)	0.010 (0.195)	0.010 (0.212)	0.009 (0.196)
Prop. size $n_j = 0.1N_j$	0.269 (0.179)	0.007 (0.179)	0.007 (0.203)	0.006 (0.182)
Non-informative				
Fixed size $n_j = 38$	0.000 (0.108)	0.000 (0.114)	0.001 (0.115)	0.001 (0.115)
Prop. size $n_j = 0.4N_j$	0.003 (0.113)	0.004 (0.120)	0.003 (0.123)	0.003 (0.122)
Fixed size $n_j = 9$	-0.007 (0.108)	-0.009 (0.115)	-0.010 (0.125)	-0.010 (0.117)
Prop. size $n_j = 0.1N_j$	-0.002 (0.110)	-0.002 (0.114)	-0.004 (0.132)	-0.003 (0.117)

Table 2

Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Second Level Standard Deviation (true value 0.632, number of replicates 1,000)

Sampling design	Unweighted estimator	Weighted estimators		
		Cluster-level weighted	Unscaled fully weighted	Scaled fully weighted
Informative at both levels				
Fixed size $n_j = 38$	0.671 (0.106)	0.638 (0.112)	0.637 (0.137)	0.604 (0.128)
Prop. size $n_j = 0.4N_j$	0.673 (0.108)	0.636 (0.112)	0.645 (0.142)	0.592 (0.130)
Fixed size $n_j = 9$	0.644 (0.145)	0.584 (0.172)	0.920 (0.289)	0.536 (0.222)
Prop. size $n_j = 0.1N_j$	0.598 (0.164)	0.546 (0.183)	1.002 (0.317)	0.498 (0.242)
Informative only at cluster level (level 2)				
Fixed size $n_j = 38$	0.595 (0.100)	0.596 (0.110)	0.605 (0.111)	0.601 (0.111)
Prop. size $n_j = 0.4N_j$	0.582 (0.096)	0.582 (0.115)	0.603 (0.113)	0.596 (0.113)
Fixed size $n_j = 9$	0.547 (0.121)	0.548 (0.135)	0.671 (0.144)	0.563 (0.133)
Prop. size $n_j = 0.1N_j$	0.538 (0.122)	0.535 (0.142)	0.696 (0.158)	0.551 (0.139)
Non-informative				
Fixed size $n_j = 38$	0.611 (0.086)	0.612 (0.092)	0.621 (0.090)	0.617 (0.091)
Prop. size $n_j = 0.4N_j$	0.609 (0.084)	0.606 (0.088)	0.626 (0.088)	0.618 (0.088)
Fixed size $n_j = 9$	0.561 (0.105)	0.561 (0.112)	0.685 (0.119)	0.575 (0.111)
Prop. size $n_j = 0.1N_j$	0.551 (0.109)	0.546 (0.113)	0.703 (0.134)	0.559 (0.112)

Estimation of ω_σ .

The results in Table 2, concerning ω_σ , are more difficult to interpret (Table 2: rows 1-8, columns 2-4). First note that also in the base scenario the estimation of ω_σ is biased, especially in small samples. Therefore the weight-based adjustment should be judged as effective if it is able to reproduce the same bias which is observed in the base

scenario. On these grounds the behavior of the scaled fully weighted estimator is satisfactory in nearly all situations, with the exception of the small samples when the design is informative at both levels. In that case there is also a not negligible number of replications which yielded a zero estimate for ω_σ (4.5% for the design with fixed size and 2% for the design with proportional size). The unscaled fully

weighted estimator does not suffer from the problem of null estimates, but, apart from having a larger variance than the scaled version, tends to overestimate ω_{σ} , showing a relative bias of about 50% in small samples when the design is informative at both levels. Note also that the scaled fully weighted estimator outperforms the cluster-level weighted estimator even when the design is informative only at level 2.

4.2.4 Additional Simulations Using the Model with Covariates

Some additional simulations were conducted to assess the performance of the scaled fully weighted estimator in the model with one cluster-level covariate and one subject-level covariate. The model is the same used in the main simulation plan, except for the inclusion of a covariate at each hierarchical level. For each covariate the values are generated from a standard Gaussian distribution, while the corresponding regression coefficient is fixed to 0.1.

As shown by Tables 3 and 4, the scaled fully weighted estimator is effective in removing the bias induced by the informative design. Relative to the unweighted estimator the sampling variance is higher, especially for the subject-level regression coefficient. Overall, the performance of the weighted estimator is satisfactory.

Table 3

Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Regression Coefficient of the Subject-Level Covariate (true value 0.1, number of replicates 1,000)

Sampling design	Non informative	Informative at both levels	
	Unweighted estimator	Unweighted estimator	Scaled fully weighted estimator
Fixed size $n_j = 38$	0.101 (0.028)	0.117 (0.040)	0.098 (0.050)
Prop. size $n_j = 0.4N_j$	0.099 (0.026)	0.117 (0.043)	0.098 (0.052)
Fixed size $n_j = 9$	0.099 (0.055)	0.119 (0.083)	0.100 (0.104)
Prop. size $n_j = 0.1N_j$	0.098 (0.056)	0.116 (0.089)	0.098 (0.107)

Table 4

Simulation Means and Standard Deviations (in parenthesis) of Point Estimators of the Regression Coefficient of the Cluster-Level Covariate (true value 0.1, number of replicates 1,000)

Sampling design	Non informative	Informative at both levels	
	Unweighted estimator	Unweighted estimator	Scaled fully weighted estimator
Fixed size $n_j = 38$	0.096 (0.119)	0.117 (0.130)	0.102 (0.142)
Prop. size $n_j = 0.4N_j$	0.102 (0.110)	0.106 (0.133)	0.106 (0.142)
Fixed size $n_j = 9$	0.094 (0.117)	0.116 (0.141)	0.105 (0.150)
Prop. size $n_j = 0.1N_j$	0.094 (0.119)	0.115 (0.144)	0.095 (0.158)

4.2.5 General Remarks

Our simulations showed that the PML approach is, in most cases, a simple and effective strategy to deal with informative sampling designs. The only requirement is the knowledge of the inclusion probabilities at every stage of the sampling process (except when the informativeness does not concern all the levels).

As for the regression parameters, the scaled version of the fully weighted estimator showed good performance in our simulations, achieving a low bias with a modest increase in the sampling variance (in some cases the variance even diminished). Even when weighting is superfluous, the loss of efficiency due to the inclusion of scaled weights is very low.

While for the estimation of the regression parameters weighting seems to be always effective, for the variance component ω_{σ} attention should be paid to the sample size: in fact, weighting leads to satisfactory results only when the cluster size is high, *i.e.*, when it allows a good representation of the complex variance structure. However the sample size is crucial in the estimation of ω_{σ} also when all the basic assumptions of the multilevel ordinal model are satisfied.

The differences induced by the type of clusters in the sample, fixed or variable size, are minimal, with equal sized clusters leading to slightly better estimators; however, as already noted, the important differences are largely due to the average size of the clusters in the sample.

The results of our simulation study confirm the findings of Pfeffermann *et al.* (1998) on the random intercept linear model: probability-weighted estimators are good for the intercept, while some relevant bias remains in the estimation of the variance components when the sample is small. As was to be expected, when passing from a linear to a nonlinear model the performance of the estimators slightly worsen, but the direction and importance of the bias in the various cases are similar. Also the advantages of scaling are confirmed.

The rise in the sampling variance due to the inclusion of the weights often has a magnitude which is in line with the results of Pfeffermann *et al.* (1998), though in some cases we found a reduction in the sampling variance, notably for the intercept when the weights are scaled and the design is informative at both levels. An interesting difference with respect to Pfeffermann *et al.* (1998) is the role of scaling in reducing the sampling variance: in this respect, scaling seems to be more effective in the binary model than in the linear model.

As already noted, the critical point in the random intercept binary model is the estimation of the cluster-level variance ω_{σ} , which represents a difficult task also when the

design is non-informative. Using the threshold formulation outlined in section 2, ω_{σ} is defined as ω/σ , so estimation of ω_{σ} involves the problems observed in the linear model associated with estimation of the two variance components. The simulations showed that the performance of the scaled weighted estimator of ω_{σ} is not entirely satisfactory in the case of small sample sizes. A possible way to improve the performance of the estimator is the adoption of a different scaling method. Korn and Graubard (2003) investigated the issue of scaling in the context of the linear model and warned that the scaling method here adopted ('scaling method 2' of Pfeffermann *et al.* 1998) may be badly biased under some designs, even if the sample size of clusters and sample sizes within the clusters are large. To get an idea of the extent of the bias we performed a short simulation study under the unfavorable scenario outlined by Korn and Graubard (2003), namely a simple random sample of clusters whose population sizes are all equal, and a simple random sample of individuals within each sampled cluster that is of size $2m$ or $m/2$ for a fixed m , depending on whether the observed variability of the individuals within the clusters tends to be large or small, respectively. In this case the scaled weights at subject level are all equal to 1, so weighting becomes ineffective. As a consequence, in the linear variance component model the within variance will be biased high. To see how this behavior extends to the random intercept binary model we simulated 1,000 datasets with 80 clusters and cluster sizes of 36 or 9 depending on whether the binomial variance of the responses of the cluster is over or under the median, respectively. Under the same superpopulation model as in the main simulations, the simulation means (and standard deviations) are -0.003 (0.098) for β_{σ} and 0.451 (0.144) for ω_{σ} . The cluster-level variance is heavily underestimated, though its value is not so far from the worst case of the main simulations (0.498 under the informative design with $n_j = 0.1N_j$). Therefore, it seems unlikely to encounter situations where the bias is much greater than already shown by our simulations. Obviously, if estimation of the variance components is of primary interest it is important to improve the method, but this requires further research.

4.2.6 Bootstrap Variance Estimation

The estimated covariance matrix of the parameter estimates obtained by inversion of the information matrix, yielded by default by the NLMIXED procedure, is not reliable when using the weighted estimators to adjust for an informative design. For example, the estimated standard error of the scaled fully weighted estimator under the design informative at both levels with $n_j = 0.4N_j$ is 0.109 for β_{σ} (compared with a Monte Carlo value of 0.183) and 0.089 for ω_{σ} (compared with a Monte Carlo value of 0.130). For

the other sampling sizes similar downward biases arise, so an alternative variance estimator should be devised.

The bootstrap procedure described in section 3.4 has been applied to estimate the sampling standard deviations of the weighted point estimators of β_{σ} and ω_{σ} . We limited the analysis to the scaled fully weighted estimator and to designs that are informative at both levels. To save computational resources we implemented a bootstrap procedure which omits the steps concerning the elementary units, *i.e.*, only the clusters are resampled. This procedure is expected to produce sufficiently accurate results, given the low sampling fraction (35/300) of the clusters (see section 3.4). Each simulation comprises 1,000 replications. For every replication the values of the response variable are generated through the two-stage scheme described in section 4.1 and 200 bootstrap samples are selected. Table 5 reports, for each parameter, the Monte Carlo standard error of the sampling distribution of the scaled weighted estimator on 1,000 replications of the complex design (see Tables 1 and 2), the corresponding average bootstrap estimate and the relative bias.

Table 5
Simulation Standard Deviations of the Scaled Weighted Point Estimators of the Intercept and of the Second Level Standard Deviation and Corresponding Bootstrap Estimates (with 200 Bootstrap Samples Each) for Designs Informative at Both Levels (1,000 Replicates for Each Design)

Sampling design Inform. Both levels	β_{σ}			ω_{σ}		
	Simul. s.d.	Boot. Estim.	Relative error	Simul. s.d.	Boot. Estim.	Relative error
Fixed size $n_j = 38$	0.185	0.175	-5.4%	0.124	0.106	-14.5%
Prop. size $n_j = 0.4N_j$	0.183	0.173	-5.5%	0.140	0.129	-7.9%
Fixed size $n_j = 9$	0.200	0.167	-16.5%	0.234	0.599	156.0%
Prop. size $n_j = 0.1N_j$	0.195	0.173	-11.3%	0.247	0.538	117.8%

Due to the extremely long computational time, we limited our experiment to a specific bootstrap procedure based on only 200 bootstrap samples. Further work is needed to calibrate the number of bootstrap samples and to explore possible variants of the method. Nonetheless, the entries of Table 5 give some hints about the behavior of bootstrap estimators.

The performance is better for the estimation of the sampling standard deviation of the estimator of β_{σ} , rather than of ω_{σ} . Especially for ω_{σ} the sample size is the critical factor: for small cluster sizes ($n_j = 9$ and $n_j = 0.1N_j$) the bootstrap estimate is completely unreliable. On the contrary with large cluster sizes ($n_j = 38$ and $n_j = 0.4N_j$) the results are quite good, since for both β_{σ} and ω_{σ} the bootstrap produces a slight underestimation of the true variance. Note, however, that the bad performance of the variance

estimator for ω_{σ} is not as critical since Wald tests for variance parameters are not generally recommended in ordinary situations anyway.

5. FINAL REMARKS

The wide use of multilevel ordinal and binary models in many fields of application has motivated our study on the effects of complex sampling designs on the fitting of such models. In the paper we showed, by means of simulations, the bias induced by a two-stage complex sampling design on the fitting of a simple random intercept binary model when the clusters and/or the subjects are selected with probabilities that depend on the model's random terms. The simulation study also showed that in such situations the bias can be reduced in an effective manner by the probability-weighted estimation procedure (PML) described in the paper, which is easily implemented in the SAS environment. In particular, the scaled version of the weighted estimator achieved, for both fixed and random parameters, a low bias with a modest increase in the sampling variance. Even when weighting is superfluous, the loss of efficiency due to the inclusion of scaled weights seems to be very low.

The application of the proposed methodology to real life examples requires an operational strategy which depends on the extent of the available information on the sampling design. Two extreme cases can be envisaged: a) for each stage of the sampling plan, the probabilities of inclusion and the adjustments for poststratification and nonresponse are exactly known; b) the information is limited to the final overall weights, which also include adjustments for poststratification and nonresponse.

In case a) the weights can be calculated at each sampling stage as the reciprocals of the product of sample selection probabilities and response probabilities given the sample selection, with a further correction for a possible poststratification. This is the idea behind the real life application presented in Pfefferman *et al.* (1998).

In case b) the lack of information is critical, since, even in the absence of nonresponse and poststratification, it is not possible to disentangle the cluster-level and the (conditional) subject-level weights, at least without strong assumptions. As a result, weighted estimation cannot be performed.

Between the two extreme cases just outlined there are many possible intermediate situations which require *ad hoc* solutions. For example, a common case arises when the researcher has access to the cluster-level inclusion probabilities (π_j) and to the final overall subject-level weights (w_{ij}), which also include adjustments for poststratification and nonresponse. When the poststratification and

nonresponse affect only the subject level, then the subject-level (conditional) weights can be calculated as $w_{ij}^* = w_{ij} \cdot \pi_j$. Another more complex situation is described by Korn and Graubard (2003).

A drawback of probability-weighted estimation is the need for special procedures to estimate the variability of the estimators. In our application we adopted a bootstrap technique, which is conceptually simple and easy to program, but requires some computational effort. Our limited simulation study suggests that its performance is good only for large sample cluster sizes; however more simulations would be needed to fully understand the behavior of the bootstrap estimator.

Another open question is the choice of the most effective scaling method for reducing the bias of the estimator of the variance components when the sample size is small.

The PML approach described in the paper is absolutely general and the estimation technique based on the NLMIXED procedure of SAS is easy to generalize to other nonlinear models. Therefore it would be of interest to assess the performance of the method in models other than the random intercept binary model here considered.

APPENDIX A

We report the SAS code used for implementing the probability-weighted (PML) estimators described in the paper. The essential part of the code is the NLMIXED procedure of SAS, which is a general procedure for fitting nonlinear random effects models using adaptive Gaussian quadrature. Though the NLMIXED procedure does not include an option for PML estimation, it is still possible to insert the weights in the likelihood, using different tricks for level 1 and level 2 weights. To insert level 1 weights it is necessary to exploit the option which allows to write down the expression for the conditional likelihood of the model: then one should simply translate in SAS programming statements the expression $w_{ij} \log L_{ij}(\theta|u)$ (see section 3.1). On the other hand, level 2 weights can be inserted in the likelihood through the `replicate` statement. Unfortunately, this statement is limited to integer weights, so to avoid gross approximations it is advisable to proceed as follows: a) inflate all the level 2 weights by an arbitrary constant k (equal to 10,000 in our application); b) insert the integer part of the inflated weights in the likelihood through the `replicate` statement; c) multiply the estimated covariance matrix by k by means of the `cfactor` option. This trick relies on the fact that multiplying the level 2 weights by a constant has the only effect of inflating the information matrix by that constant, leaving the estimates unchanged. Anyway, when using the weighted estimation

method to adjust for an informative design the estimated covariance matrix of the parameter estimates is not reliable.

In the following the SAS code is reported, where the symbols `/*` and `*/` include the comments:

```
proc nlmixed data=dataname qpoints=10
cfactor=10,000;
/* cfactor is a constant multiplying the
estimated covariance matrix of the parameter
estimates */
parms b0=0 sd=0.5; /* initial values */
bounds sd >= 0;
eta=b0+randeff*sd;
if (yobs=1) then z=probnorm(eta);
else if (yobs=0) then z=1-probnorm(eta);
if (z >= 1e-8) then ll=log(z); else ll=-1e100;
/*to avoid numerical problems if z becomes
too small*/
ll=ll*w1_2; /* inclusion of level 1 weights
*/
model yobs~general(ll);
random randeff ~normal(0,1) subject=j;
/* j is the cluster identifier */
replicate w2; /* inclusion of level 2
weights (only integers) */
ods output ParameterEstimates=pe
ConvergenceStatus=cs;
run;
```

ACKNOWLEDGEMENTS

We wish to thank two anonymous referees for their suggestions which contributed to substantially improve the paper.

REFERENCES

- GIBBONS, R.D., and HEDEKER, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*. 53, 1527-1537.
- GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*. 73, 43-56.
- GRAUBARD, B.I., and KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*. 5, 263-281.
- HEDEKER, D., and GIBBONS, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*. 50, 933-944.
- KORN, E.L., and GRAUBARD, B.I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society B*. 65(1), 175-190.
- PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*. 61(2), 317-337.
- PFEFFERMANN, D., SKINNER, C.J., HOLMES, D.J., GOLDSTEIN, H. and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*. 60(1), 23-40.
- RASBASH, J., BROWNE, W., GOLDSTEIN, H., YANG, M., LEWIS, I., HEALY, M., WOODHOUSE, G. and DRAPER, D. (1999). A users guide to MLwiN. London: Multilevel models project, Institute of Education, University of London.
- RENARD, D., and MOLENBERGHS, G. (2002). Multilevel modeling of complex survey data. In *Topics in Modelling of Clustered Data* (Ed. M. Aerts M). London: Chapman and Hall. 263-272.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.
- SAS INSTITUTE (1999). SAS/STAT User's Guide Version 8. Cary: SAS Institute Inc.
- SHAO, J., and TU, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer-Verlag.
- SINGH, A.C., FOLSOM, R.E. and VAISH, A.K. (2002). A hierarchical Bayes generalization of the Fay-Herriot method to unit level nonlinear mixed models for small area estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. New York, August 10-13. 3258-3263.
- SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). Chichester: Wiley. 59-87.

Longitudinal Analysis of Labour Force Survey Data

GEOFF ROWE and HUAN NGUYEN¹

ABSTRACT

The Canadian Labour Force Survey (LFS) was not designed to be a longitudinal survey. However, given that respondent households typically remain in the sample for six consecutive months, it is possible to reconstruct six-month fragments of longitudinal data from the monthly records of household members. Such longitudinal micro-data – altogether consisting of millions of person-months of individual and family level data – is useful for analyses of monthly labour market dynamics over relatively long periods of time, 25 years and more.

We make use of these data to estimate hazard functions describing transitions among the labour market states: *self-employed*, *paid employee* and *not employed*. Data on job tenure, for employed respondents, and on the date last worked, for those not employed – together with the date of survey responses – allow the construction of models that include terms reflecting seasonality and macro-economic cycles as well as the duration dependence of each type of transition. In addition, the LFS data permits spouse labour market activity and family composition variables to be included in the hazard models as time-varying covariates. The estimated hazard equations have been incorporated in the LifePaths microsimulation model. In that setting, the equations have been used to simulate lifetime employment activity from past, present and future birth cohorts. Simulation results have been validated by comparison with the age profiles of LFS employment/population ratios for the period 1976 to 2001.

KEY WORDS: Microsimulation; Censoring; Truncation; Employment dynamics.

1. INTRODUCTION

In recent years, there has been increased recognition of the importance of studying labour market dynamics using individual level (micro-) data. For this purpose, new panel surveys have been developed, for example, the Survey of Income and Labour Dynamics (SLID) (Statistics Canada 1998). But, existing LFS data (Statistics Canada 2002) provides a virtually untapped historical resource, in the form of many fragmentary event histories. From a conventional standpoint, the data currently comprises a time series of more than 300 cross-sectional surveys that were conducted monthly over more than 25 years. However, from a longitudinal perspective, those same data consist of about 6.5 million fragmentary event histories covering overlapping time intervals within the past quarter century and totalling over 34 million person-months of observation.

The analysis referred to in this paper was specifically directed towards development of hazard models to be incorporated in LifePaths (Statistics Canada 2001) – a micro-simulation model of the Canadian population. Further details on the LifePaths model are available from the Statistics Canada website at www.statcan.ca/english/spstd/index.htm.

The paper is organized in the following way. In section 2, we discuss some features of LFS data when reorganized as longitudinal records and we present three examples comparing estimates derived from the resulting longitudinal file

with corresponding estimates from other sources. In section 3, we focus on the use of the data to model employment activity for LifePaths. There, we discuss the use of LFS micro-data in estimating hazard equations that describe employment dynamics. Finally, we present some illustrations of estimation results and a validation of LifePaths simulations that make use of the hazard equations.

2. LONGITUDINAL LFS DATA: DISTINGUISHING FEATURES AND PROOF-OF-CONCEPT

A longitudinal version of the LFS data was constructed by concatenating the monthly records of individual respondents into a file containing one record per respondent. Since an LFS respondent normally remains in the LFS sample for six consecutive months, we can obtain six-month histories for most respondents. These histories are not, by themselves, long enough for most longitudinal analyses. However, given the overlapping rotation groups that are part of the LFS design, these six-month fragments may be used in analysis of the experiences of employment cohorts over decades. (In line with the focus of the analysis below, we use the term “cohort” to refer to a relatively homogeneous group for all of whom a specified initial event has occurred. Thus, an “employment cohort” might refer to all persons who started a new job within a specified time period or, more narrowly, to all of those who started their third job

¹ Geoff Rowe and Huan Nguyen, Socio-Economic Analysis and Modeling Division, Analysis and Development Branch, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

within a specified time period. The data available from the LFS determines how narrowly such a cohort can be defined here).

Figure 1, which illustrates some characteristics of the LFS data after they are formed into longitudinal records, focuses on changes in employment status for the employment cohort who started a job in January 1976. Respondents who were members of this cohort and who entered the sample through rotation 1 contribute data on the first six months, from January 1976, when the job started, to June 1976, when they left the LFS sample. For respondents from rotation 2, the six-month longitudinal data window shifts right one month (starting and ending one month later than those given by rotation 1). The overlapping data windows of respondents from subsequent rotations evolve similarly. Thus, the longitudinal LFS data can be seen as a combination of overlapping sets of panel data, in which respondents from the same rotation constitute a conventional data panel.

Successive six-month fragments of longitudinal LFS data can be combined to provide successive estimates of cumulative attrition from an initial employment cohort and, further, to identify new cohorts defined in terms either of a new job or of a period without employment. Thus, over the long term (currently up to 25 years), many different samples of individuals can contribute information about the same employment cohort observed at different points in time.

Even so, month-to-month changes are observed largely from the same sample of individuals. The two shaded areas in Figure 1 illustrate this. The respondents from each of the rotations 2-5 contribute data for both the May-June and the June-July intervals.

This is not the first attempt to use LFS data longitudinally. Stasny (1986) and Lemaître (1988) studied errors in the estimation of “gross flows” between labour force states (*employed*, *unemployed* and *not in the labour force*) over intervals of one month. Lemaître found that problems arose both because of response errors and because “Labour Force Survey concepts, designed for cross-sectional purposes, tend to “create” flows when consecutive months’ responses are linked”. (Examples include the treatment of *on-call workers* and of the *self-employed without a business*). Nevertheless, he concluded, “Administrative data have shown that not all sub-groups of status changers are seriously overestimated”. Kinack (1991) examined the longitudinal consistency of responses to questions on job search activity that were used to distinguish between the categories *unemployed* and *not in the labour force*. He found substantial inconsistency, particularly when associated with proxy responses from different proxy respondents. These studies have shown that focusing on transitions between the categories *employed* and *not employed* (i.e., without distinguishing between *unemployed* and *not in the labour force*) could help reduce the impact of response error.

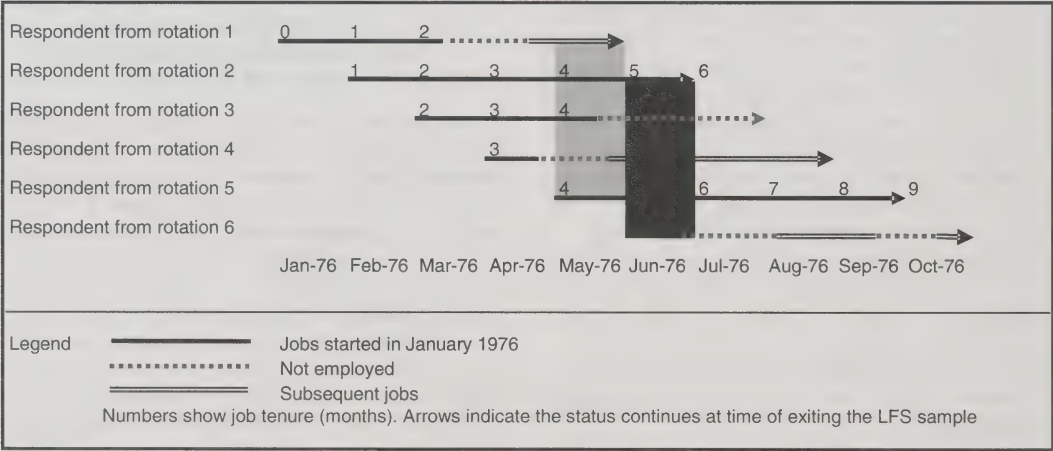


Figure 1. Illustration of LFS fragmentary data on cohort starting jobs in January 1976

Cross-sectional LFS data have previously been used to estimate frequencies of job hiring and job separation over monthly intervals (Lemaître, Picot and Murray 1992). In that case, hiring was directly observed from the frequency of reported job-tenures of one month or less, while separation was determined residually using aggregate estimates of employment change together with the estimates of hiring. Cross-sectional LFS data have also been used to calculate and compare duration statistics for synthetic-cohorts. For example, Corak and Heisz (1995) use retention rates from a single time interval to represent a hypothetical cohort's experience. Synthetic-cohort retention rates were obtained using the numbers of employed LFS respondents reporting job tenure " t " in month " m " together with those reporting tenure " $t+1$ " the next month. Such uses of cross-sectional data have certain limitations. In particular, because the movement of individuals is not directly observed, destination states are unknown. (Although we may estimate the proportion that separated from a job, we can not estimate the proportion of those that became unemployed rather than dropping out of the labour force or beginning another job immediately). Nevertheless, a time series of synthetic-cohort statistics – for example, the proportions of jobs that might last a certain duration – can serve as an index that is sensitive to changing labour market conditions.

2.1 Proof-of-Concept: Selected Examples of Longitudinal Data Validation

The LFS data were not intended to be used longitudinally and problems can arise with such use (Stasny 1986; Lemaître 1988; Kinack 1991). Consequently, it is important to verify, for each analysis individually, that valid estimates can be obtained by month-to-month comparison of

longitudinal responses. We present three examples of the verification of LFS longitudinal estimates below. In Figure 2, we compare estimates of the annual number of job separations in Canada from 1976 to 1995 (separations of all types, permanent and temporary) based on LFS data and on administrative data. The latter are based on Records of Employment (ROE) issued by employers at the time of job separation for Employment Insurance purposes (Statistics Canada 1998).

As may be seen, the number of transitions determined by month-to-month comparison of LFS data corresponds closely to the number from ROE data. Still, there are differences between the two series. Some of these differences could arise because of differences in coverage between the LFS and administrative data, as well as periodic changes in the LFS design or questionnaire. Another source of difference could arise because our counts based on LFS data neglect job separations of multiple job holders who remained employed in at least one job (*i.e.*, we counted only main-job changes). Nevertheless, we regard the degree of agreement between the LFS and administrative data as close enough to justify further analysis of the LFS micro-data. Both data sources imply that the annual rate of job separations was high: based on ROE data between 1978 and 1995, the average annual job separation rate for males was over 38 percent of annual person-jobs. Further analysis of the LFS micro-data can shed light on these dynamics.

Figure 3 goes further in the validation of employment dynamics, comparing "job survival" probabilities for males and females who started a job in 1993, as estimated from the LFS data and from SLID. (Note that 1993 corresponds to the first year of SLID data).

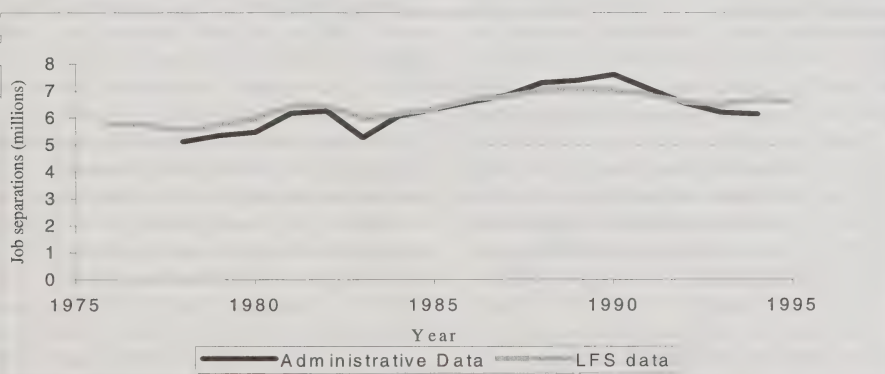


Figure 2. Estimates of Annual Job Separations

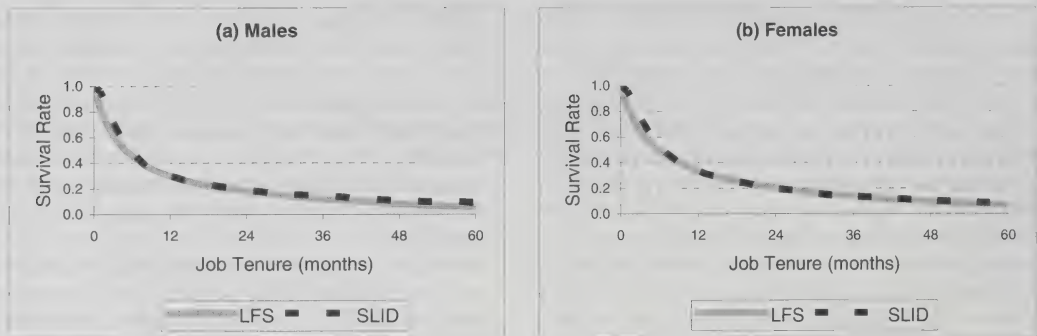


Figure 3. 'Job Survival' Probabilities of the Cohort Starting Jobs in 1993: Comparison of Estimates Based on LFS and SLID

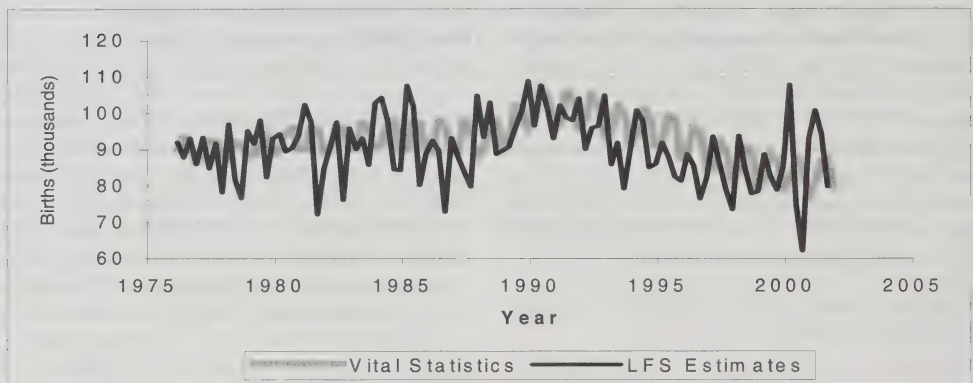


Figure 4. Estimates of Births in Canada by Quarter, 1976 - 2001

The "job survival" probabilities were estimated from LFS data by the chained product of average retention rates derived from monthly main-job separation rates over the period 1993 to 1998. Survival probabilities from the SLID data were estimated in a similar manner using the reported job tenure and dates of job end. Both survival curves display the same characteristic shape; showing relatively high attrition for jobs of duration less than a year, but with much lower attrition rates at job tenures of one to five years. There are discrepancies between the estimates for durations of about six months or less, which may be related to the one-year recall period of SLID interviews and to the restriction of LFS job-tenure data to main-jobs. However, over periods as long as five years, the LFS and SLID provide very similar estimates. And, with the available LFS data, we can track some employment cohorts for as long as 25 years after the employment spell began.

A final illustration of effective longitudinal use of LFS data involves month-to-month comparison of the number of

children aged less than one year as reported by female economic family heads or by the spouse of a male head. A infant child that is newly reported by a woman aged between 15 and 50 likely signifies the birth of a child. In order to make direct comparisons between these LFS estimates and vital statistics, we made some straightforward adjustments to account for the proportion of births occurring to other women living in economic families (*e.g.*, teen lone parents living with their parents) and for births in the Yukon, NWT and Nunavut. A comparison of the resulting LFS monthly estimates of births with the corresponding counts of births registered in vital statistics (Figure 4) demonstrates that the LFS estimates follow secular trends in fertility as well as capturing some of the month-to-month fluctuation in births. Taken together, these three examples indicate that—with careful attention to survey coverage, survey concepts and the possibility of response error—the LFS can provide useful longitudinal micro-data.

3. USING LONGITUDINAL LFS MICRO-DATA FOR MODELING EMPLOYMENT ACTIVITY IN LIFEPATHS

This section focuses on the use of the LFS data to simulate employment activity in LifePaths. Currently, LifePaths uses a 3-category classification of employment status – employee (E), self-employed (SE), and not employed (NE). We have not analyzed transitions involving unemployment. (Unemployment is a complex state requiring additional questions to ascertain and so, as noted above, unemployment transitions are particularly subject to response error).

There are six transitions that can result in a change in employment status (as represented in Figure 5). LifePaths models all of these transitions. In addition, job changes that do not appear to involve an interruption of employment are also modeled by LifePaths (denoted here as $E \Rightarrow E$). The LFS micro-data were used to estimate hazard equations for each of these seven transitions. The estimated coefficients of these equations became parameters in the LifePaths “Career Work” module. Below we discuss some technical issues that arise due to the limitations of the LFS data, followed by an illustration of the estimation results and then of a simulation outcome.

The fragmentary nature of these data poses a challenge for analysis. An important question is whether there are unavoidable biases that result from their fragmentary nature. In general, the answer is that the limitations of these data can be accounted for and potential sources of bias can be avoided with careful analysis.

3.1 Censoring and/or Truncation of Event Histories

One source of concern for an analyst of these data is the absence of retrospective employment information other than

the length of the *current* employment spell. We might think of individual employment histories as consisting of a (largely unobserved) succession of contingent employment states (illustrated in Figure 6) with transitions among these states reflecting the process of career development. Thus, given only the transitions observable within the LFS window, the transition rates that can be estimated will inevitably involve pooling data from respondents who have had markedly different prior careers. In contrast, panel surveys like SLID, collect retrospective data at the first interview that, although limited, at least permits some experience rating of respondents in terms of previous extended work interruptions or periods of part-time work.

Another concern, illustrated in Figure 6, is that LFS employment spell durations may be left-truncated and/or right-censored. Right-censoring refers to the circumstance in which a spell ceases to be observed or a respondent ceases to be at risk without a transition occurring of the type being studied. This happens either (1) because the respondent’s household “rotated out” of the LFS sample before any transition occurred, or (2) because another transition occurred that was not of the type under active study. Similarly, these data are frequently left-truncated. This refers to the circumstance in which the beginning of a spell is unobserved, because it happened before the respondent’s household “rotated in” to the LFS sample. (These data are left-truncated rather than left-censored, because respondents provide the information necessary to determine the elapsed duration of the current spell at the time of the first interview). Since both truncation and censoring are generally independent of employment event processes, neither should lead to bias in the estimation of transition probabilities, if properly accounted for in the likelihood function.

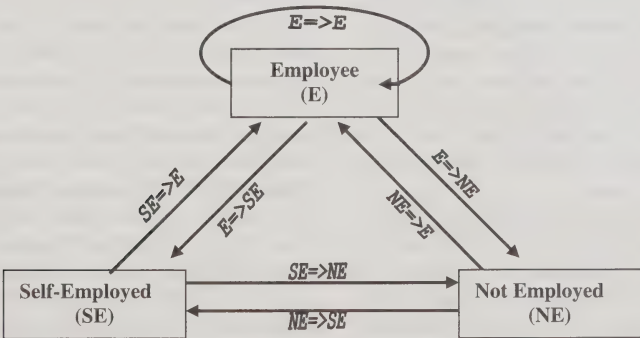


Figure 5. Employment Status and Transitions in LifePaths

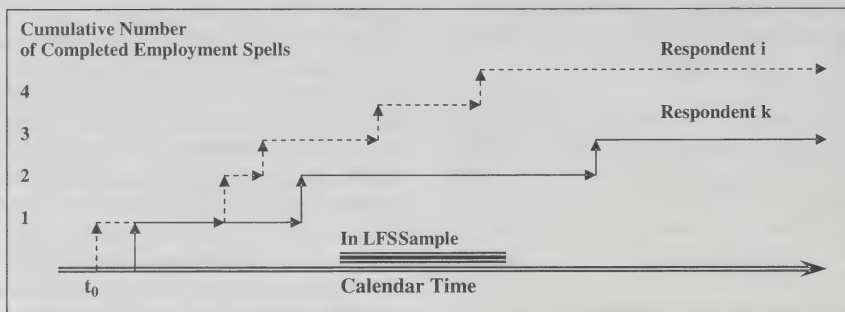


Figure 6. Recurrent Events and Employment Spell Durations Observable within the LFS Sample Window

The combination of full and partial information provided by left-truncated and right-censored data can be represented in a conditional likelihood (Wang 1991). In a competing risks framework, the likelihood of an employment transition type j involving respondent i may be expressed in terms of the spell duration observed k months after i was first observed to be at risk of transition j . Let t_i denote the year and month of the LFS interview in which i 's current employment state was first observed (i.e., often the first interview). Based on information collected at each interview, we can determine the length of the current spell of employment or spell not employed (m_{t_i}). Then $m_{t_i+k} = m_{t_i} + k$ would denote the elapsed spell duration in the state as assessed k months after the first observation – assuming no intervening events – and the likelihood of a transition of type j (i.e., L_{j,t_i+k}) can be expressed in terms of m_{t_i+k} . Terms in the likelihood function comprise: the probability density of durations leading up to transitions of type j ($f_j(m_{t_i+k})$), the corresponding cumulative probability ($F_j(m_{t_i+k})$), a binary variable indicating whether or not censoring has occurred (C_{j,t_i+k}), and a further binary variable indicating whether or not the current spell was left-truncated ($LT_{i,j}$). Note that, in the competing risks framework, the density $f_j(m_{t_i+k})$ relates to a latent variable – the waiting time leading specifically to transition j – and that we must assume there is one such density for each competing event. In principle, the completed spell duration (observed when a transition occurs) will correspond to the minimum of competing, latent waiting times.

To account for left truncation, the likelihood is expressed in terms of conditional probabilities given the spell duration first observed (m_{t_i}): these probabilities take the form either of conditional probabilities evaluated at the time of an observed transition ($f_j(m_{t_i+k}|m_{t_i})$) or of conditional probabilities of surviving – without the occurrence specifically of transition j – to the observed duration

($1 - F_j(m_{t_i+k}|m_{t_i})$), depending on whether or not censoring has occurred.

$$L_{i,j,t_i+k} = f_j(m_{t_i+k}|m_{t_i})^{1-C_{j,t_i+k}} (1 - F_j(m_{t_i+k}|m_{t_i}))^{C_{j,t_i+k}} \\ = \frac{f_j(m_{t_i+k})^{1-C_{j,t_i+k}} (1 - F_j(m_{t_i+k}))^{C_{j,t_i+k}}}{(1 - F_j(m_{t_i}))^{LT_{i,j}}} \quad (1)$$

This likelihood accounts for all of the information we have regarding the specific risk of transition j and can incorporate the effect of other competing risks by treating them as censoring events that are in addition to censorship by “rotating out” of the sample. Competing risks problems are commonly formulated in terms of such latent waiting times, especially in epidemiology and biostatistics, but also in economics (e.g., Heckman and Honoré 1989). However, while providing a mathematically convenient motivation for the likelihood, the approach has been criticized “on the basis of unwarranted assumptions, lack of physical interpretation and identifiability problems” (Prentice, Kalbfleisch, Peterson, Floumoy, Farewell and Breslow 1978).

The conditional likelihood (1) can be approximated by a Poisson likelihood (Holford 1980; Laird and Olivier 1981), thereby also acknowledging the discreteness of the data (i.e., transitions are generally “observed” in the one month interval between successive interviews). Equation (1) can be re-expressed in terms of a binary variable (Y_{j,t_i+k}) that represents occurrence or non-occurrence of a transition in a particular time interval (note that $Y_{j,t_i+k} = 1 - C_{j,t_i+k}$). Then, Y_{j,t_i+k} is treated as a Poisson random variable having an expected value equal to the hazard “ h_{j,t_i+k} ” which is assumed piecewise constant. Under this model, the contribution from i to the log-likelihood over n periods (using $h_{j,t_i+k} = f_j(m_{t_i+k}) / (1 - F_j(m_{t_i+k})) = -\partial \ln(1 - F_j(m_{t_i+k})) / \partial m_{t_i+k}$ together with (1)) is approximately:

$$\ln(L_{i,j}) \approx \sum_{k=1}^n \left[Y_{j,t_i+k} \ln(\hat{h}_{j,t_i+k}) - \hat{h}_{j,t_i+k} \right]. \quad (2)$$

It is common practice to account for a complex survey design by means of a “pseudo” likelihood that incorporates the survey weight. Maximizing the “pseudo” likelihood corresponds to minimization of a weighted sum of deviance terms (*i.e.*, terms representing the difference between estimated likelihood contributions and their maximum possible values). Thus, the full-sample, conditional log-likelihood for transition j may be transformed into a weighted deviance D_j (note that W is derived from the survey weights and, since transitions are typically identified by comparing employment states between interviews, we use averages of consecutive cross-sectional survey weights to obtain W):

$$D_j \approx -2 \left(\sum_i \left[\sum_{k=1}^{n_i} W_{t_i+k} Y_{j,t_i+k} \ln(\hat{h}_{j,t_i+k}) \right] + \sum_i \left[\sum_{k=1}^{n_i} W_{t_i+k} \left[Y_{j,t_i+k} - \hat{h}_{j,t_i+k} \right] \right] \right). \quad (3)$$

In the analysis of each transition type j , we treat other events (*i.e.*, non- j events occurring to the same population-at-risk) as censoring, and so the deviance for a set of such events will be the sum of component deviances (*i.e.*, if the overall hazard is the sum of competing hazards, then the competing risks may be treated as independent (Prentice *et al.* 1978)).

A more direct motivation of the same deviance takes Poisson processes as its starting point (Borgan 1984; Andersen 1985; Andersen and Borgan 1985; Lawless 1987), rather than starting with postulated event-specific, latent, duration densities like $f_j(m_{t_i+k})$. In this case, we can model sampled multivariate counting processes that represent the number of occurrences of each specific transition in time intervals $[t_0, t)$. Sample counting processes, represented by the step functions in Figure 6, are observable counterparts of cumulative hazard functions. The assumption that the underlying hazard functions are approximately piecewise constant leads directly to the Poisson deviance as an approximation (Lindsey 1995). To limit bias, the principal concerns are that the population-at-risk can be identified, that censoring or truncation mechanisms are conditionally independent of the underlying employment processes and that the intervals over which hazards are assumed constant are not too large.

It is possible to obtain simple averaged estimates of employment hazard functions (such as those displayed in Figure 3) by implicitly splicing together all available information on members of a defined cohort from the

longitudinal LFS samples. (That is, maximizing likelihood (1), but without considering any covariates). Making allowance for censoring and truncation in this way is a relatively simple example of such problems compared with the more complex observation schemes considered by Alioum and Commenges (1996). This implicit splicing of information is apparent in the deviance (3) which has two components: the first component is non-zero only at observed transitions, while the second component reflects the weighted differences between cumulative events and cumulative hazards (accumulated over all durations prior to the events or to censoring times). To the extent that the LFS cross-sections are representative samples for each reference week, then – taken together – they will provide an accurate estimate of the numbers of events occurring over the “life” of an employment cohort. Similarly, within samples from employment cohorts, we can expect to find left-truncated and right-censored respondent spells that might fill-in the missing prior histories of those left-truncated spells that terminate with a transition. As such, the first component of the deviance will accurately reflect whether hazard estimates tend to be large over periods where observed events are frequent. And the second component, summed over all respondent-months, may have a value similar to that which we might have obtained had there been no left-truncation. So, for data as extensive as these, the conditional likelihood may be almost equivalent to an unconditional likelihood.

3.2 Estimating Employment Transition Hazard Equations

Patterns of employment transition differ significantly among different demographic groups. For example, full-time students are most active in the labour market during their summer break, whereas the maternity leave that an employed pregnant woman takes may be largely determined by Employment Insurance regulations. Accordingly, LifePaths distinguishes among the following groups and models their employment activities separately:

- Those who are full-time students;
- Those who have just graduated or left school and are in a transition to an after-school job;
- Pregnant women for whom a maternity-leave may apply;
- Those who are in prime ages of employment; and
- Older workers in transition to retirement.

We discuss here only the estimation for the fourth group, comprising individuals who are in what is referred to in LifePaths as their “career employment” phase (the most important phase in terms of impact on the economy). Particulars for the other groups are available from the Statistics Canada website noted above.

For implementation in LifePaths, our hazard model uses a log-linear form of regression equation – one equation for each of the 7 transitions and for each sex separately, giving a total of 14 equations:

$$E(Y_{j,t_1+k}) \approx \hat{h}_{j,t_1+k} = \exp \left(\hat{g}(m_{j,t_1+k}) + X_{j,t_1+k} \hat{\beta} \right) \quad (4)$$

where $E(\cdot)$ is the expectation operator, $g(m)$ is a log-linear spell duration spline, X is a vector of time-varying covariates and β is a vector of regression coefficients. The term $g(m)$ corresponds to a piecewise Weibull baseline hazard, which, in our specification, distinguishes employment transition risks at durations of less than a year from risks at durations of more than a year. The covariates, X , include variables representing individual age, education, province of residence, presence of children by age group, spouse's employment status, calendar month and calendar

year, as well as interactions among some of these factors. Final estimates of β and $g(m)$ minimize the deviance (3).

The only example of detailed results that we present here involves the mutual influence of husband's and wife's employment status on each other's respective transition hazards. Figure 7 compares coefficient estimates from the seven equations that correspond to the seven transitions we specified. The two panels correspond to the separate sets of equations for males and females. The category "no spouse present" was treated as the reference category and the spouse's employment status was classified into "with paid employment", "self-employed", and "not employed". The estimated coefficients are presented here in terms of risk relative to the reference group. Thus, with other covariates controlled, the hazard of becoming self-employed for female employees whose husbands are self-employed is about 2.5 times higher than the hazard of their counterparts who do not have a spouse (see tallest bar in the top panel).

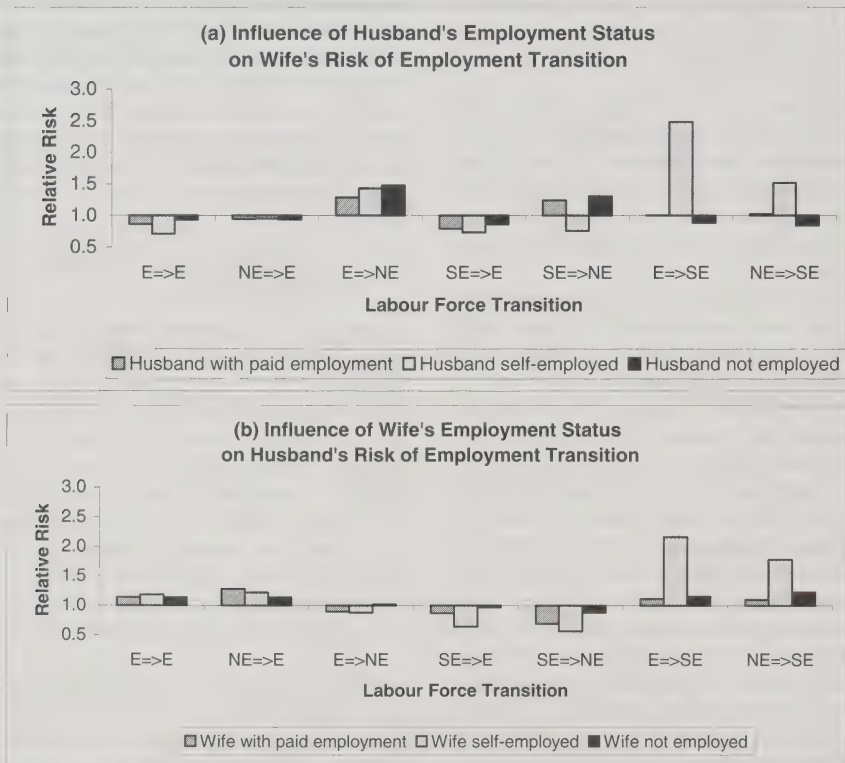


Figure 7. Impact of Spouse's Employment Status on Employment Transition Risks

Figure 7 shows that the very presence of a spouse can work in opposite directions for males and females. The most frequent transitions for both sexes are $E \Rightarrow E$, $NE \Rightarrow E$ and $E \Rightarrow NE$. For females, the first two of those transitions are less likely to occur to married women than to single women, while the transition to "not employed" is more likely. (The presence of children is not the reason for this, as their presence is accounted for by other terms in the equation). For males, the pattern is reversed. Thus, these results appear consistent with conventional gender roles. However, taking account of the magnitudes of these relative risks, we are not given the impression that gender roles have a particularly strong influence after the influence of other variables is credited.

Figure 7 reveals another conspicuous pattern. First, the relative risks of a transition into self-employment, for spouses with husbands/wives in self-employment, stand out as the highest among all other transitions. In addition, spouses with husbands/wives in self-employment have the lowest relative risks of a transition out of self-employment. Thus, self-employment status seems to be mutually reinforcing within families. These observations are consistent with forms of joint self-employment involving a family business (*e.g.*, a corner store) or involving endogamy among professionals (*e.g.*, lawyers marrying other lawyers).

4. FROM ESTIMATED PARAMETERS TO THE SIMULATION RESULTS: AN ILLUSTRATION

Our example of the role of spouse's employment status points to the need for family context in the simulation of employment activities. It is a challenge for LifePaths to

integrate these relationships into the simulation process. For example, if individual education progression or the effects of education on employment transitions are not modeled appropriately and accurately, then the consequences will cascade from direct education-employment relationships to a chain of indirect impacts, involving relationships between education and marriage, fertility, interprovincial migration, *etc.* These impacts would then spill over to the simulated spouse, as indicated above. It is not difficult to see that, unless these relationships are specified appropriately and the parameters are estimated with reasonable accuracy, bias would be spread over a wide range of simulated outcomes.

An overall validation of the LifePaths employment hazard equations was obtained by comparing simulated annual average employment/population ratios with direct cross-sectional estimates from the LFS. The simulated employment/population ratios were obtained from a synthetic population whose members were exposed appropriately to one or other of the seven types of employment hazards over the course of each simulated year. The simulated employment/population ratios were calculated from the resulting annual person-years of employment in the synthetic population: that is, these ratios are an outcome of simulated flows into and out of employment. The simulations necessarily involved generating appropriate distributions of covariates that in turn determine the distributions of employment transition hazards. As may be seen in Figure 8, LifePaths accurately reflects the age patterns of female employment in both 1976 and 2001 and correspondingly accounts for the dramatic change observed in those age patterns over the past quarter century.

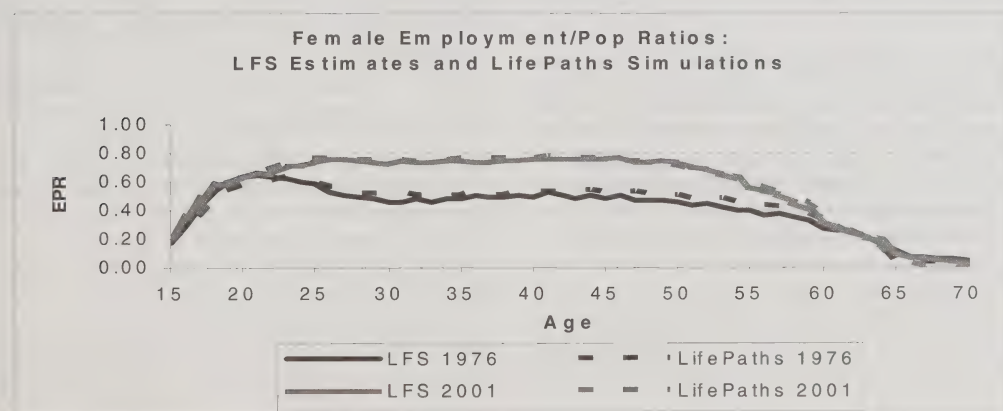


Figure 8. Validating hazard equations using LifePaths

5. CONCLUSIONS

We have demonstrated that the LFS data – when organized into the fragmentary event histories collected over the six-month periods that most respondents spend in the sample – represents a significant longitudinal micro-data asset. There is sufficient sample and breadth of content to provide for important analysis of labour market dynamics and, conceivably, of demographic processes such as fertility. Moreover, the data is monthly and spans more than a quarter century, so that analysis based on it has uninterrupted time depth that is unique in Canada.

In our main application (employment transitions), other results (not reported here) appear to confirm the influence of a range of explanatory variables on an individual's chances of an employment transition. These covariates include age, job tenure (or duration not employed), educational attainment, presence of young children (especially for women), province of residence, seasonality, and business cycles. However, this work is still in its initial stage and, to date, our approach to inference has been informal. Future work will involve extending and refining our models and establishing a more rigorous basis for evaluation of the models.

REFERENCES

- ALIOUM, A., and COMMENGES, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, 52, 512-524.
- ANDERSEN, P. (1985). Statistical models for longitudinal labor market data based on counting processes. In *Longitudinal Analysis of Labor Market Data*. (Eds. James J. Heckman and Burton Singer). Cambridge University Press, Cambridge.
- ANDERSEN, P., and BORGAN, Ø. (1985). Counting process models for life history data: A review. *Scandinavian Journal of Statistics*, 12, 97-158.
- BORGAN, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics*, 11, 1-16.
- CORAK, M., and HEISZ, A. (1995). The duration of unemployment: A user guide. Research Paper Series No. 84, Analytical Studies Branch, Statistics Canada.
- HECKMAN, J.J., and HONORÉ, BO.E. (1989). The identifiability of the competing risks model. *Biometrika*, 76(2), 325-330.
- HOLFORD, T.R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics*, 36, 299-305.
- KINACK, M. (1991). Measuring data quality with longitudinal data. 1991 *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 514-519.
- LAIRD, N., and OLIVIER, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76, 231-240.
- LAWLESS, J.F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82, 808-815.
- LEMAÎTRE, G. (1988). The measurement and analysis of gross flows. Working Paper, Labour and Household Surveys Analysis Division, Statistics Canada.
- LEMAÎTRE, G., PICOT, G. and MURRAY, S. (1992). Workers on the move: An overview of labour turnover. *Perspectives on Labour and Income*, 4(2), Statistics Canada.
- LINDSEY, J.K. (1995). Fitting parametric counting processes by using log-linear models. *Applied Statistics*, 44, 201-212.
- PRENTICE, R.L., KALBFLEISCH, J.D., PETERSON, A.V., FLOURNOY, JR. N., FAREWELL, V.T. and BRESLOW, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34, 541-554.
- STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.
- STATISTICS CANADA (1998). Permanent layoffs, quits and hirings in the Canadian Economy: 1978-1995. Catalogue # 71-539-XIB.
- STATISTICS CANADA (1998). Survey of Income and Labour Dynamics: A Survey Overview. Catalogue # 75F0011XPB, <http://www.statcan.ca/english/freepub/75F0011XIE/free.htm>.
- STATISTICS CANADA (2001). The LifePaths Microsimulation Model: An Overview. <http://statcan.ca/english/spsd/LifePaths.htm>.
- STATISTICS CANADA (2002). Guide to the Labour Force Survey. Catalogue # 71-543-GIE, <http://www.statcan.ca/english/IPS/Data/71-543-GIE.htm>.
- WANG, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86, 130-143.

Contact and Cooperation in the Belgian Fertility and Family Survey

MARC CALLENS and CHRISTOPHE CROUX¹

ABSTRACT

Combining response data from the Belgian Fertility and Family Survey with individual level and municipality level data from the 1991 Census for both nonrespondents and respondents, multilevel logistic regression models for contact and cooperation propensity are estimated. The covariates introduced are a selection of indirect features, all out of the researchers' direct control. Contrary to previous research, Socio Economic Status is found to be positively related to cooperation. Another unexpected result is the absence of any considerable impact of ecological correlates such as urbanity.

KEY WORDS: Nonresponse; Multilevel analysis; Fertility and Family Survey.

1. INTRODUCTION

The aim of this paper is to empirically assess the relative importance of correlates of contact and cooperation rates in the Belgian Fertility and Family Survey (FFS Belgium 1991).

The conceptual and theoretical nonresponse framework used in this paper has been proposed by Groves and Couper (G&C 1998). In their view, nonresponse arising from noncontact is directly influenced by survey design features such as the number and the timing of calls. Conditionally on these survey design features, other important features such as physical impediments of the housing units and accessible-at-home patterns of the would-be respondents, which are indirectly measured by various social environmental and socio-demographic attributes, also play an important role. The decision to cooperate or to refuse is primarily regarded as a direct function of a dynamic social communicative process between the interviewer and the interviewee. Survey design, main interviewer, sample person and social environment characteristics are considered to have only an indirect influence on cooperation rates.

We use both individual level and municipality level data from the 1991 Census data, matched to the fieldwork outcome variable for nonrespondents and respondents of the 1991 Belgian FFS. In this survey, individuals are the sampling units. It is a face-to-face survey with low noncontact (4%) and moderate refusal rates (22%). We consider our data to be hierarchically nested with sample units at the lower and municipalities at the higher level. Including covariates at both levels, multilevel logistic regression models for contact and cooperation propensity

are estimated. The covariates are a selection of indirect features, all out of the researchers' direct control.

Some intriguing results are: (1) Socio Economic Status indicators like education are positively related to cooperation and (2) ecological factors including urbanicity are not correlated with nonresponse. This is in contrast with findings from previous US-based research.

2. A THEORY FOR CONTACTABILITY AND COOPERATION

The process of realising an interview consists of two major components: the process of contacting a sample person and dependent on contact, the process of cooperation with a survey request. An attractive multi-level theoretical framework for studying contactability and cooperation has been proposed by Groves and Couper (G&C 1998).

2.1 Contactability

Chronologically, the process of contacting a sample person comes first. Some sample persons are never contacted by interviewers and hence never make a decision about their survey cooperation. Relative to the process of cooperation, the process of contacting a sample person is quite simple.

G&C (1998) consider contactability to be a function of three factors: (1) whether there are any physical impediments that prevent interviewers to get in touch with the sample person, (2) when sample persons are at home and (3) when and how many times the interviewer tries to contact the sample person. The number and timing of calls

¹ Marc Callens, Population and Family Study Centre (CBGS), Markiesstraat 1, B-1000 Brussels, Belgium & Dept. of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail: Marc.Callens@cbgs.be; Christophe Croux, Dept. of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail: Christophe.Croux@econ.kuleuven.ac.be.

by the interviewer and the accessible-at-home patterns of the sample persons are the proximate causes of contactability. The accessible-at-home patterns of the sample person are affected by the presence of physical impediments (*e.g.*, telephone presence), socio-demographic attributes (*e.g.*, commuting times) and social environmental attributes (*e.g.*, crime). Also survey design features such as the length of the data collection period and the interviewer workload might have an influence on contact rates.

2.2 Cooperation

The central question in the survey stage following contact is why sample persons do or do not cooperate with the interviewer request. In the Groves-Couper model to study cooperation, the proximate causes of the decision to cooperate or to refuse lie at the level of the household and his or her interaction with the interviewer. Another component in the theoretical framework of G&C (1998) is the set of survey design features, such as: the agency of data collection, advance warning of the survey request, topic saliency, *etc.*

G&C (1998) consider also two factors that are out of the control of the survey designer: influences of the sample person and social environmental influences. These variables are not considered to be direct causal influences on cooperation, but indirect measures of what are essentially social psychological constructs. Important theoretical constructs in this respect are: opportunity costs, social exchange and social isolation.

2.2.1 Opportunity Costs

The notion of opportunity costs implies that sample persons weigh the opportunity costs in agreeing to spend their time responding to a survey interview. An important ingredient in the opportunity costs theory is the amount of discretionary time for the sample person available to complete the survey. Those with less discretionary time are less likely to feel free to participate in a survey. Some indirect indicators for the amount of discretionary time are: the inverse of the number of adults in a household and (the amount) of labour force participation. Of course, there are also obligations away from employment tasks such as commitments to friends and relatives that also might raise the opportunity costs of a survey.

2.2.2 Social Exchange

Social exchange theory considers the perceived value of equity of long-term associations between persons or between a person and societal institutions (Blau 1964). Central to all conceptualisations of social exchange is the notion that, unlike economic exchange, all social commodities are part of an intuitive bookkeeping system in which debts (*e.g.*, obligations) and credits (*e.g.*, expectations) are

taken into account (G&C 1998). The social exchange perspective can be applied whenever there is an ongoing relationship between the survey organisation and the sample person (*e.g.*, government surveys).

Those receiving fewer services from the government may – in considering the cumulative effect of multiple government contacts – feel less need to cooperate. Since government services are disproportional across socio-economic strata, indicators of Socio-Economic Status (SES) should reflect exchange influences on survey participation. However, a major problem with social exchange theory is that two alternative hypotheses between SES and cooperation might be deduced from it (G&C 1998). First, one can argue that lower SES groups may have the greatest indebtedness to the government for the public assistance they may receive. Higher SES groups feel far less that they owe any sort of repayment. In this perspective, the relationship between socio-economic status and cooperation propensity is a negative one. Alternatively, a curvilinear relationship between SES and cooperation may be hypothesised. The lowest SES groups may believe that they are disadvantaged routinely compared to more fortunate people. The highest SES groups feel themselves repeatedly targeted in terms of time and money but receive little in return. In such a hypothesis, both the highest and the lowest SES feel relatively deprived in the relationship with large-scale social institutions and tend to refuse survey cooperation.

2.2.3 Social Isolation

Closely related to the social exchange hypothesis is the social isolation hypothesis. Social isolates are out of touch with the mainstream culture of a society: they tend to behave in accordance with subcultural norms or in explicit rejection of those of the dominant culture. They are believed to be less likely to participate in a variety of social and political activities, including responding to surveys (Couper, Singer and Kulka 1997). In terms of SES, social isolation theory implies a positive relationship between SES and cooperation: lower SES groups are resentful of their dependence on the government, whereas higher SES groups have a greater sense of civic obligation. Such a positive relationship between SES and social isolation is opposite to the relationships predicted by social exchange theory.

Demographic indicators of social isolation are race, ethnicity, age and gender; with minorities, elderly and men in the role of the relatively isolated. Indicators of social isolation at the micro-level include whether the sample person lives in a single-person household, whether the sample person has any children, whether the sample person has moved recently and whether the sample person lives in a large multiunit structure.

2.2.4 Urbanicity

At the community level contextual factors such as urbanicity, population density, crime rates and lack of social cohesion are hypothesised to influence survey cooperation. Residents of rural areas tend to cooperate at a higher level compared to residents in towns. However, it is not clear which mechanism is responsible for this urbanicity effect which might be explained in terms of greater population density, higher crime rates and higher social disorganisation that are associated with life in urban areas. Population density is hypothesised to reduce cooperation through the experience of crowding. Fear of crime may produce an unwillingness to provide information to strangers. Finally, urban life is associated with social disorganisation, characterised by weakened local kinship and friendship networks and reduced participation in local affairs.

3. DATA AND METHOD

3.1 Data

In this study we make use of both aggregated and micro-level data of the Belgian 1991 Census linked to the response status for respondents and nonrespondents from the Belgian Fertility and Family Survey (FFS-Belgium 1991) held shortly after the Census operations.

3.1.1 The FFS Survey (1991)

The Fertility and Family Survey in Belgium was organised by the Population and Family Study Centre (CBGS), a Scientific Institute from the Flemish Government. This survey was carried out between April and October 1991, which is very close to the decennial census date: April 1 in the same year. The main focus of the FFS-project is on reproductive behaviour, to be seen however in the broader context of partnership and family history, and the interaction between employment and reproduction (Cliquet and Callens 1993; Callens 1995). The target population consists of men and women of Belgian nationality, born in the period 1951-1970 and with main residence in the Flemish Region of Belgium.

A two-stage cluster sampling design was used for men and women separately. In a first stage, municipalities were selected from various socio-economic strata (Vanneste 1989). In each selected municipality, individuals were selected at random. In this way 2,975 women and 1,989 men were selected to take part in the survey. A fieldwork method was used to compensate for non-response: stratified random substitution of nonrespondents of the target sample by persons selected from a reserve sample (Chapman 1983; Vehovar 1999).

The final sample size, *i.e.*, including the substitution operation, equals 4,776 persons (2,897 women and 1,879 men). In this study we make use of respondents and nonrespondent cases of both the initial target sample and the fieldwork substitution operation ($N = 6,847$).

Among both men and women, the nonresponse can be ascribed in 7 out of 10 cases to a refusal to participate in the survey. In 2 out of 10 cases, nonresponse is due to the fact that the persons selected could not be contacted, and in 1 out of 10 cases, an interview was impossible because of sickness, language difficulties or some other reason.

3.1.2 Matching 1991 Census Person-Level Data (1991)

Our primary source of information on both respondent and nonrespondent cases is provided by the 1991 Census.

In an effort to reconcile privacy concerns and scientific interests, we used a simple technique to make the matching of person-level Census data and survey data anonymous. We provided a dataset to the National Institute of Statistics (NIS) containing only the national identification number and the response status for each respondent and nonrespondent case. As a result of the matching operation by the NIS, we received a selection of the 1991 Census data enriched with only two survey variables: the response status variable and an indicator whether a sample person belongs to the base or substitute sample.

The 1991 Census individual level data we have at our disposal are: the individual form and the house unit form. The individual form contains information about: the place of residence, the nationality, the labour force activity status, the first marriage, the birth year of the children, education and professional activities. The house unit form includes information on the housing unit of the household such as: the type of housing unit, the number of housing units in the building, ownership, building period, the number of rooms and corresponding squared meters, the presence of a telephone and comfort indicators such as the number of bath rooms.

3.1.3 Contactability and its Determinants

To study the process of contactability, we ideally need data on the outcomes of all successive attempts to contact sample persons. In this study however, we do not have such detailed information at our disposal: we only know the final outcome of each survey request. Therefore, we can only study the probability of ever making contact with the sample person (coded 1 = contact and coded 0 = non-contact) and not whether it was easy or difficult to make contact. Sample persons that are known not to reside (anymore) on the sample address we do consider contacted. At 241 out of 6,847 sample units (3.52%), all contact attempts failed.

The data we use are measured at two levels: the individual level ($n = 6,847$) and the municipality level ($n = 123$). At the sample person level, we consider three types of variables: physical impediments to contact sample persons, reasons for sample persons to be present in their homes and control variables.

As there are no direct interviewer observations of physical impediments available to us, we have to rely solely on indicators for physical impediments available in the Census data. Three variables are used: whether the housing unit is a single-family structure or not, whether the housing unit is large (more than 10 units) or not and whether the sample person has a telephone or not.

Determinants of at-home patterns in this study are: civil status (unmarried, married and divorced), age (20-24, 25-29, 30-34 and 35-39 years) and activity status (inactive vs. other). For women only, we also consider the number of children (0, 1, 2 and 3+). For those in the labour force we have also detailed information about: working part-time vs. working full-time, the number of weekly working hours (<21, 21-35, 36-42, >42 hours), employment status (employee vs. own-account), having a second job or not and working at home or not.

We also use two control variables: substitution (whether a sample person originates from the base target sample or from the substitution sample) and gender (whether a sample person comes from the female sample or from the male sample).

At the municipality level ($n = 123$), we use five variables: population density (persons per square km for the residence of the sample person), urban status (the cities of Antwerp and Gent vs. other municipalities), percentage multi-unit structures (in quartile format: <7.13, 7.13-15.14, 15.14-27 and >27), percentage homes owner-occupied (in quartile format: <64.5, 64.5-71, 71-77.7 and >77.7) and percentage persons of minority race (in quartile format: <0.90, 0.9-2.22, 2.22-5.29 and >5.29).

3.1.4 Cooperation and its Determinants

We are interested in the probability of ever getting cooperation (coded 1 = cooperation and coded 0 = non-cooperation) conditionally on contact; not whether it was easy or difficult to get cooperation from the sample person. For 1,399 out of 6,606 contacted sample persons (21.18%), all attempts to get cooperation failed.

Again, the data we use are measured at two levels: the individual level and the municipality level. At the sample person level, we have indicators for the opportunity costs hypothesis, the exchange hypothesis and the isolation hypothesis. Substitution is used as a control variable.

Indicators for the opportunity costs hypothesis are: activity status (inactive vs. other), working part-time vs.

working full-time, the number of weekly working hours (<21, 21-35, 36-42, >42 hours) and employment status (employee vs. own-account).

Indicators for Socio-Economic Status in our study are: the surface of the living rooms (in squared meters: <65, 65-84, 85-104, 105-124 and >125), the number of bathrooms (0, 1 and 2+) and educational level (primary, secondary – first stage, secondary – second stage, high – non-university and high – university level). Other exchange hypothesis indicators are: whether one receives a replacement income from the government or not and whether the house is owner-occupied or not.

Indicators for the social isolation hypothesis are: gender, civil status (unmarried, married and divorced), age (20-24, 25-29, 30-34 and 35-39 years), single-family structure of the housing unit and for women only: the number of children (0, 1, 2 and 3+) and the presence of children under the age of five years. Finally, substitution is included as a control variable.

At the municipality level, we use the same five variables as in section 3.1.3: urban status, population density, percentage multiunit structures, percentage owner-occupied and percentage persons of minority race.

3.2 Method of Analysis

3.2.1 Bivariate χ^2 -Test

In a first exploratory series of analyses of the correlates of contactability and cooperation, we calculate percentages for two-way contingency tables and include the results for the χ^2 -test of independence against association. Such a χ^2 -test, like any significance test, indicates the degree of evidence for the existence of an association, not the strength of an association. When at least one variable is ordinal, more powerful tests of independence than the χ^2 -test such as the linear trend test do exist, but for reasons of simplicity of presentation, we do not use them in this paper.

3.2.2 Multilevel Logistic Regression

In a second series of analyses, we use multilevel logistic regression to simultaneously estimate the impact of the various determinants (Snijders and Boskers 1999). We opt for the use of a multilevel method, because we regard our data as hierarchically nested with individuals at the lower level (level 1) and municipalities at the higher level (level 2).

Let p_{ij} be the probability that an individual i belonging to municipality j is contacted (or cooperates). We will consider four different models for explaining this probability: the null random model, two versions of the random intercept model and the standard logistic regression model.

The empty or unconditional model does not take explanatory variables into account. We specify the model such that

logit transformed probabilities p_{ij} have a normal distribution:

$$\text{logit}(p_{ij}) = 1/(1 + \exp(p_{ij})) = \gamma_0 + u_{0j}$$

where γ_0 is the population average and u_{0j} the random deviation from this average for group j . These deviations u_{0j} are assumed to be independent normally distributed random variables with mean zero and variance τ_0^2 .

When there are r variables at the individual level that are potentially explicative for the observed outcomes, then they are incorporated as a linear function in the random intercept model:

$$\text{logit}(p_{ij}) = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + u_{0j}$$

where $\gamma_1, \dots, \gamma_r$ are the slope parameters measuring the effect of the explicative variables.

If we would drop the random effects u_{0j} then we obtain a standard logistic regression model:

$$\text{logit}(p_{ij}) = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij}$$

By also including s variables at the community level, we get an intercept model with both level-1 and level-2 covariates:

$$\text{logit}(p_{ij}) = \gamma_0 + \sum_{h=1}^r \gamma_h x_{hij} + \sum_{k=1}^s \gamma_k x_{kij} + u_{0j}$$

We use SAS Proc Nlmixed (SAS Institute 1999) to actually estimate the parameters. In SAS Proc Nlmixed an adaptive version of Gauss-Hermite Quadrature (numerical integration) is used to solve the maximum likelihood estimation problem. To test if a specific parameter equals zero, a Likelihood Ratio χ^2 -test is used.

4. RESULTS

4.1 Contactability

Table 1 presents the bivariate results by the χ^2 -test of the percentage never contacted by various indicators of physical impediments. One strong correlate is whether the housing unit is a single-family structure or not, the latter having much higher noncontact rates (8.1%) than other units (2.4%). Also, sample persons living in large multiunit housing structures tend to have higher noncontact rates (11%) than those not living in large multiunit housing structures (3.1%). Another strong correlate is the presence of a telephone: 9.7% of those with no telephone were never contacted.

Table 1
Percentage Never-Contacted by 'Physical Impediments' Attributes

Physical impediments attributes	Percentage never contacted	χ^2	df	p
Single-Family Structure		97.6	1	<0.0001
No	8.1			
Yes	2.4			
Large multi-unit structure (>10)		38.4	1	<0.0001
No	3.1			
Yes	11.0			
Telephone		88.9	1	<0.0001
No	9.7			
Yes	2.7			

Table 2 shows the bivariate results for contactability by 'reasons to be present at home' attributes. Relatively more unmarried (4.4%) and divorced (6.9%) sample persons than married (2.9%) sample persons are never contacted. There are much lower rates of noncontacts among those that are inactive (0.9%) compared to other persons (3.5%). Having at least 3 or more children (0.9%) leads to low noncontact rates, compared to having two children (2.6%) or at most 1 child (4%). Those working at home (1.5%) and those being an independent worker (1.9%) show modestly lower noncontact rates than those working elsewhere (3.6%) or those working as an employee respectively (3.6%). Age, the number of weekly working hours, working part-time vs. full-time and having a second job or not have no significant influence on contactability.

Table 2
Percentage Never-Contacted by 'Reasons to be Present at Home' Attributes

Reasons to be present at home	Percentage never contacted	χ^2	df	p
Civil status		19.4	2	<0.0001
Unmarried	4.4			
Married	2.9			
Divorced	6.9			
Inactive vs. other		4.0	1	0.04
Inactive	0.9			
Other	3.5			
Number of children ^a		14.5	3	0.0023
0	4.3			
1	4.0			
2	2.6			
3+	0.9			
Employment place ^b		4.6	1	0.03
At home	1.5			
Elsewhere	3.6			
Employment status ^b		4.0	1	0.05
Employee	3.6			
Own-account	1.9			

^a subsample of women only (n=4,098)

^b subsample of active persons only (n=5,368)

In addition, substitution is associated with higher noncontact rates (5.9%) compared to the base sample (2.6%). No significant difference has been found for the male and the female subsample.

In a multiple logistic regression model of the combined effects of those individual-level indicators that have some marginal bivariate effect on contactability only single-family structure ($\chi^2 = 35.75$, $p = <0.0001$), telephone ($\chi^2 = 52.63$, $p = <0.0001$) and substitution ($\chi^2 = 28.59$, $p = <0.0001$) remain significant.

In Table 3, noncontact rates for various environmental attributes are presented. Cities (6.6%) have higher non-contact rates compared to nonurban areas (3.1%). The percentage never contacted is higher for high-density areas (5.4%) than low-density areas (1.7%). The presence of multiunit structures and the presence of persons of other nationalities tend to increase non-contact rates. Finally, the percentage of owner-occupied houses shows a negative association with noncontact rates.

Table 3

Percentage Never-Contacted by 'Environmental' Attributes

Environmental attribute	Percentage never contacted	χ^2	df	p
Urban status		24.0	1	<0.0001
Cities	6.6			
Other	3.1			
Population density		34.4	3	<0.0001
Lowest quartile	1.7			
Second quartile	3.2			
Third quartile	3.8			
Highest quartile	5.4			
% Multi-unit structures		50.4	3	<0.0001
Lowest quartile	2.0			
Second quartile	2.2			
Third quartile	4.0			
Highest quartile	5.9			
% Persons of other nationalities		23.1	3	<0.0001
Lowest quartile	2.5			
Second quartile	2.3			
Third quartile	4.3			
Highest quartile	4.8			
% Homes owner-occupied		64.4	3	<0.0001
Lowest quartile	6.4			
Second quartile	3.6			
Third quartile	1.6			
Highest quartile	2.7			

We complement now the bivariate analysis with a multivariate analysis. In Table 4 four models for modelling contact relative to noncontact are presented. Model 1 is the null random model at the municipality level. Model 2 is a multiple logistic regression model. In this model, we have

included the person-level effects that remained significant in a multivariate context (*i.e.*, single-family structure, telephone and substitution) and the variable activity status because of its theoretical importance. Model 3 is a random intercept version of model 2. In Model 4, we have extended Model 3 with the municipality level variable multi-units structures (in %) only.

Table 4

Results of (Multilevel) Logistic Regression Models of Contactability

Results	Model 1: Null Random	Model 2: Logistic Regression	Model 3: Random Intercept Level 1	Model 4: Random Intercept Level 1 & 2
<i>Intercept</i>	4.01*** (0.16)	3.08*** (0.73)	3.68*** (0.77)	4.15*** (0.79)
<i>Individual Characteristics</i>				
Single-family structure		1.16*** (0.15)	1.02*** (0.17)	0.92*** (0.17)
Telephone		1.19*** (0.16)	1.25*** (0.17)	1.26*** (0.17)
Inactive vs. other		-1.23 (0.72)	-1.34 (0.75)	-1.33 (0.74)
Substitution sample		-0.78*** (0.14)	-0.64*** (0.15)	-0.62*** (0.15)
<i>Municipality Characteristics</i>				
Multi-unit structures (%)				-0.02* (0.01)
<i>Estimated variances</i>				
Var(Intercept)	1.03		0.82	0.79
<i>Goodness of fit</i>				
Deviance	1,720	1,658	1,606	1,599

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, one-tailed tests.

The effects of the person-level covariates in Models 2, 3 and 4 are in accordance with the findings of the bivariate analysis. Single-family structure and the presence of a telephone have a positive influence on contactability, while the effect of activity status is not significant. The impact of field substitution is negative. We also notice a (rather small) reduction of the regression coefficient for single-family structure and substitution in the multilevel models 3 and 4. Models 3 and 4 have one variance component for the intercept. To test the null hypothesis that the random intercept variance equals zero, we use the Likelihood Ratio test and compare the conventional logit model (Model 2) with the random intercept model (Model 3). The difference in deviance between both models is large (52). So, there might be some variance in the intercept to explain by municipality level covariates. By introducing municipality characteristics one at a time, we can test for significant effects by calculating deviance differences between Model

4 and Model 3. The only deviance difference of importance noted is the case of the variable ‘multi-unit structures’ (7 units difference). No differences in deviances are found for the introduction of the other level-two variables (urban status, percentage owner occupied, population density and persons of other nationalities).

We consider Model 3 and Model 4 as the better models. According to these multilevel models, noncontact rates vary considerably across municipalities. However, the municipality level covariates in our study are not able to explain much of this variation.

4.2 Cooperation

In Table 5, we present the bivariate results for the opportunity costs hypothesis indicators. Being inactive or not does not seem to have an effect on the cooperation rate. However, when we use indicators of discretionary time, such as working part-time versus working full-time or the weekly number of working hours, the predicted negative relationship does show up in the bivariate results. In addition, self-employed sample persons have lower cooperation rates compared to employees.

Table 5
Percentage Cooperation by ‘Opportunity Cost Hypothesis’ Indicators

Opportunity cost indicators	Percentage cooperated	χ^2	df	p
Inactive vs. other		0.41	1	0.52
Inactive	77.0			
Other	78.9			
Part-time vs. Full-time ^b		10.04	1	0.001
Part-time	82.3			
Full-time	77.4			
Number of working hours ^b		15.3	3	0.0016
<20	80.1			
21-35	84.7			
36-42	77.6			
> 43	75.7			
Employment status ^b		4.2	1	0.04
Employee	78.7			
Own-account	74.6			

^b subsample of active persons only (n=5,180)

The predictions of the exchange hypothesis theory do not show up in the bivariate results presented in Table 6. SES indicators like the surface of the living room and the number of bathrooms are not negatively, but positively related to cooperation. Of course, these measures are not ideal, because we are not able to control for household size. Another indication of a positive relationship between cooperation and SES is the case of educational level.

Whether one receives a replacement income or not and whether the house is owner-occupied or not has no impact on cooperation rates.

In a multiple logistic regression model of the combined effects of those social exchange indicators that have some marginal bivariate effect on cooperation, only the effects of educational level ($\chi^2=39.35$, $df=4$, $p<0.0001$) and surface of the living room ($\chi^2=13.4$, $df=4$, $p=0.0095$) remain significant.

Table 6
Percentage Cooperation by ‘Exchange Hypothesis’ Indicators

Exchange indicators	Percentage cooperated	χ^2	df	p
Surface living rooms (m ²)		26.8	4	<0.0001
< 65	74.8			
65 - 84	77.6			
85 - 104	78.6			
105 - 124	79.9			
> 125	83.1			
Number of bathrooms		7.9	2	0.02
0	74.2			
1	78.6			
2	83.5			
Educational level		46.7	4	<0.0001
Primary	76.6			
Secondary, first stage	74.5			
Secondary, second stage	78.7			
High, non-university	85.1			
High, university	82.2			
Replacement income		0.3	1	0.58
No	78.7			
Yes	79.5			
Owner occupied		3.4	1	0.06
No	77.4			
Yes	79.4			

In the section for the exchange hypotheses, we have found support for the notion that those with low SES, cooperate less with surveys than those in the high SES groups. Such a positive relationship between SES is predicted by the social isolation hypothesis. Demographic indicators of social isolation theory are gender, civil status and age (See Table 7). No effects are found for gender, civil status (however, divorced sample persons are probably less cooperative) and single-family structure. Age seems to have a negative effect on cooperation. For women only, we have also data on the presence of children. We find that the number of children has a positive effect on cooperation rates. The age of the children is also important: the presence of young children is associated with higher cooperation.

The control variable substitution has a slightly negative effect on cooperation ($\chi^2=4.24$, $p=0.039$) with lower

cooperation rates for the substitution sample (77.3%), compared to the base sample (79.5%).

Table 7

Percentage Cooperation by 'Social isolation Hypothesis' Indicators

Social isolation indicators	Percentage cooperated	χ^2	df	p
Gender		1.56	1	0.21
Male	78.1			
Female	79.3			
Civil status		3.11	2	0.21
Unmarried	79.8			
Married	78.6			
Divorced	75.4			
Single-family structure		0.76	1	0.38
No	78.9			
Yes	77.7			
Age		17.5	3	0.0006
20 - 24	80.8			
25 - 29	80.7			
30 - 34	78.3			
35 - 39	75.5			
Number of children ^a		18.2	3	0.0004
0	77.9			
1	76.3			
2	81.7			
3+	84.9			
Presence of young children ^a		12.3	1	0.0005
No	77.8			
Yes	82.8			

^a subsample of women only ($n=3,955$)

Table 8 contains the bivariate results for social environmental differences in cooperation. Population density has a curvilinear effect on cooperation. Being a resident in a large metropolitan area has no effect. Thus, the evidence for the literature that crowding and high levels of stimulus input are negatively associated with cooperation is of a mixed nature.

The effect of indicators for social cohesion is not clear. Only the variable percentage owner-occupied has a (curvilinear) effect. The variables percentage persons of other nationalities and percentage multi-unit structures seem to have no effect.

Finally, we present in Table 9 a series of regression models for cooperation similar to those in section 4.1. In these models, we have included four individual level covariates: surface of the living room (<84 , >84 m²), education (up to secondary -second stage vs. high level), age (20-29, 30-39 years) and substitution sample. Surface of the living room and education have been selected as the only significant exchange hypothesis indicators in the previously described multiple logistic regression model.

Age was the only significant effect in the bivariate analysis on the social isolation hypothesis. Finally, substitution is introduced to control for possible fieldwork effects. The slightly negative effect of substitution in Model 2 might indicate that fieldwork substitution negatively influences cooperation. However, this effect disappears completely when a random intercept is introduced (Models 3 and 4). The effects of the other individual level covariates are in accordance with the findings of the bivariate analysis and do not change across Models 2 to 4. SES indicators like education and surface of the living room have a positive effect and age has a negative effect on cooperation. These effects rather confirm the social isolation hypothesis than the exchange hypothesis.

Table 8

Percentage Cooperation by 'Environmental' Attributes

Environmental attribute	Percentage cooperated	χ^2	df	p
Urban status		0.84	1	0.36
Cities	80.1			
Other	78.7			
Population density		10.7	3	0.014
Lowest quartile	80.0			
Second quartile	79.9			
Third quartile	76.0			
Highest quartile	79.4			
% Multiunit structures		3.1	3	0.38
Lowest quartile	80.1			
Second quartile	79.2			
Third quartile	77.9			
Highest quartile	78.1			
% Homes owner-occupied		12.3	3	0.0063
Lowest quartile	79.7			
Second quartile	76.2			
Third quartile	78.5			
Highest quartile	80.9			
% Persons of other nationalities		5.2	3	0.16
Lowest quartile	77.7			
Second quartile	77.6			
Third quartile	79.6			
Highest quartile	80.2			

The only level two variable of (modest) importance is multi-unit structures (in %) and has been kept in Model 4. The Likelihood Ratio test for introducing this variable gives a difference of two units in deviance terms. The introduction of one or more other second level variables gives Likelihood Ratio tests differences close to zero in deviance terms. We consider Model 3 and 4 as the most suitable models. The difference in deviance terms between model 3 and model 2 is 8 units, which is significant. The variance for the intercept term is moderate (0.21). The introduction of second level covariates (including multi-unit structures)

leaves this variance term practically unchanged. Therefore, we may state that environmental attributes like urbanicity are not important for explaining cooperation.

Table 9

Results of (Multilevel) Logistic Regression Models of Cooperation

Results	Model 1: Null Random	Model 2: Logistic Regression	Model 3: Random Intercept Level 1	Model 4: Random Intercepts Level 1&2
<i>Intercept</i>	1.41*** (0.06)	1.24*** (0.06)	1.30*** (0.08)	1.39*** (0.10)
<i>Individual Characteristics</i>				
Substitution sample		-0.15* (0.07)	-0.03 (0.07)	-0.02 (0.07)
Surface living rooms		0.23*** (0.06)	0.24*** (0.06)	0.24*** (0.06)
Educational level		0.45*** (0.08)	0.47*** (0.08)	0.47*** (0.08)
Age		-0.23*** (0.06)	-0.23*** (0.06)	-0.23*** (0.06)
<i>Municipality Characteristics</i>				
Multi-unit structures (%)				-0.006 (0.004)
<i>Estimated variances</i>				
Var(Intercept)	0.21		0.21	0.21
<i>Goodness of fit</i>				
Deviance	6,664	6,664	6,596	6,594

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, one-tailed tests.

5. DISCUSSION

In this paper, we have used 1991 individual and municipality level Census data matched to the response status variable of the Belgian Fertility and Family Survey to analyse the relative importance of correlates of contact and cooperation.

We have organised our analysis according to the Groves-Couper conceptual framework. In the bivariate analysis stage, we have found essentially the same kind of correlates as was predicted and actually found in an US-based multi-survey analysis (G&C 1998). One important difference between the present study and the US-results seems to be the nature of the effect of SES indicators (e.g., education) on cooperation. In the present study, we find a positive relationship; in the US-study the inverse relationship is found. We can imagine two alternative explanations for these conflicting findings. A first one is based on survey design effects such as topic saliency. The FFS-survey in Belgium might be atypical in being disproportionally attractive to the higher educated

because of the specific content of the survey. Replicating the present analysis for surveys about varying topics can easily test such a hypothesis. Another possible hypothesis is that effects of education on survey cooperation do vary across societies. Then the challenge is to find out why this relationship varies across countries. Such a hypothesis is far less easy to test in real, as data for several countries are needed.

In the multilevel logistic regression analysis stage, the impact of all but one contextual factor completely vanished. Only the impact of the variable percentage of multi-unit structures shows, however only weakly, some resistance against ecological randomness present in the random intercept models. To us, this is a very intriguing result. Random ecological variation at the municipality level seems to dominate largely even the urban-rural dichotomy. A possible explanation is that the variation at the community level is dominated by interviewer effects, not by ecological factors.

ACKNOWLEDGEMENTS

This paper grew out of our activities in the design and organisation of surveys at the Population and Family Study Centre (CBGS), a scientific Institute of the Flemish Government in Belgium. We are grateful to the Belgian National Statistical Institute, the local authorities and the National Register for their cooperation. We also thank the Editor and two anonymous referees for their valuable comments and suggestions. Finally, this research was supported by a FWO grant Bijzondere doctoraatsbeurs 2002-2003.

REFERENCES

- BLAU, P.M. (1964). *Exchange and Power in Social Life*. New York: John Wiley & Sons, Inc.
- CALLENS, M. (1995). *De 'Fertility and Family Survey' in Vlaanderen (Nego V, 1991), De gegevensverzameling*. Brussel: CBGS-document. 1995, 4.
- CHAPMAN, D.W. (1983). The impact of substitution on survey estimates. In *Incomplete Data in Sample Surveys, Proceedings of the Symposium*, (Eds. W.G. Madow and I. Olkin). New York: Academic. 2, 45-61.
- CLIQUET, R.L., and CALLENS, M. (Ed.) (1993). *Gezinsvorming in Vlaanderen. Hoe en Wanneer?* Brussel: CBGS Monografie 1.
- COUPER, M.P., SINGER, E. and KULKA, R.A. (1997). Participation in the decennial census: Politics, privacy, pressures. *American Politics Quarterly*. 26, 59-80.
- GROVES, R.M., and COUPER, M. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.

- SAS INSTITUTE INC. (1999). *SAS/STAT® User's Guide*. Version 8, Cary, NC: SAS Institute Inc.
- SNIJDERS, T.A.B., and BOSKERS, R.J. (1999). *Multilevel Analysis, an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications.
- VANNESTE, D. (1989). *Economische Typering Van de Belgische Gemeenten*. Leuven: Leuvense Geografische Papers.
- VEHOVAR, V. (1999). Field substitution and unit nonresponse, *Journal of Official Statistics*. 15, 2, 335-350.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 20, No. 1, 2004

The Effect of Multiple Weighting Steps on Variance Estimation Richard Valliant	1
Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations Martin H. Félix-Medina and Steven K. Thompson	19
Nonresponse in Time: Some Lessons from Major Finnish Surveys Kari Djerf	39
Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands Marianne Houbiers	55
A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers Jan Pickery and Geert Loosveldt	77
Frequency Reports Across Age Groups Bärbel Knäuper, Norbert Schwarz, and Denise Park	91
Extracting Confidential Information from Public Documents: The 2000 Department of Justice Report on the Federal Use of the Death Penalty in the United States David J. Algranati and Joseph B. Kadane	97
American Fact Finder: Disclosure Limitation for the Advanced Query System Sam Hawala, Laura Zayatz, and Sandra Rowland	115
Keys to Successful Implementation of Continuous Quality Improvement in a Statistical Agency David A. Marker and David R. Morganstein	125
In Other Journals	137
Corrigendum	139

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Volume 31, No. 4, December/Décembre 2003, 355-490

New Editor for <i>The Canadian Journal of Statistics</i>	355
Un nouveau rédacteur en chef pour <i>La revue canadienne de statistique</i>	356
Belkacem ABDOUS, Kilani GHOUDI & Bruno RÉMILLARD: Nonparametric weighted symmetry tests	357
Sharon L. LOHR & N.G.N. PRASAD: Small-area estimation with auxiliary survey data	383
Peilin SHI, Jane J. YE & Julie ZHOU: Minimax robust designs for misspecified regression models	397
Brajendra C. SUTRADHAR & R. Prabhakar RAO: On quasi-likelihood inference in generalized linear mixed models with two components of dispersion	415
Denis LAROCQUE: An affine-invariant multivariate sign test for cluster correlated data	437
Zhide FANG: Extrapolation designs with constraints	457
Giovanni PETRIS & Luca TARDELLA: A geometric approach to transdimensional Markov chain Monte Carlo	469
Correction: Yann GUÉDON & Christiane COCOZZA-THIVENT: Nonparametric estimation of renewal processes from count data	483
Forthcoming papers/Articles à paraître	484
Index: Volume 31 (2003)	485
Volume 32 (2004): Subscription rates	489
Frais d'abonnement	490

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférentiellement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ; o, O, 0; l, 1).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).
5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Volume 31, No. 4, December/Décembre 2003, 355-490

355	New Editor for <i>The Canadian Journal of Statistics</i>
356	Un nouveau rédacteur en chef pour <i>La revue canadienne de statistique</i>
357	Belkacem ABDOUS, Kilani GHODI & Bruno RÉMILLARD: Nonparametric weighted symmetry tests
383	Sharon L. LOHR & N.G.N. PRASAD: Small-area estimation with auxiliary survey data
397	PeiLin SHI, Jane T. YE & Julie ZHOU: Minimax robust designs for misspecified regression models
415	Brajendra C. SUTRADHAR & R. Prabhakar RAO: On quasi-likelihood inference in generalized linear mixed models with two components of dispersion
437	Denis LAROCQUE: An affine-invariant multivariate sign test for cluster correlated data
457	Zhide FANG: Extrapolation designs with constraints
469	Giovanni PETRIS & Luca TARDELLA: A geometric approach to transdimensional Markov chain Monte Carlo
483	Correction: Yann GUÉDON & Christiane COCOZZA-THIVENT: Nonparametric estimation of renewal processes from count data
484	Forthcoming papers/Articles à paraître
485	Index: Volume 31 (2003)
489	Volume 32 (2004): Subscription rates
490	Frais d'abonnement

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 20, No. 1, 2004

The Effect of Multiple Weighing Steps on Variance Estimation	Richard Valliant	1
Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations	Martin H. Felix-Medina and Steven K. Thompson	19
Nonresponse in Time: Some Lessons from Major Finnish Surveys	Kari Djerf	39
Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands	Marianne Houbiers	55
A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviews	Jan Pickery and Geert Loosveldt	77
Frequency Reports Across Age Groups	Bärbel Knäuper, Norbert Schwarz, and Denise Park	91
Extracting Confidential Information from Public Documents: The 2000 Department of Justice Report on the Federal Use of the Death Penalty in the United States	David J. Algerant and Joseph B. Kadane	97
American Fact Finder: Disclosure Limitation for the Advanced Query System	Sam Hawala, Laura Zayatz, and Sandra Rowland	115
Keys to Successful Implementation of Continuous Quality Improvement in a Statistical Agency	David A. Marker and David R. Morganstein	125
In Other Journals		137
Corrigendum		139

All inquiries about submissions and subscriptions should be directed to jos@sch.se

BIBLIOGRAPHIE

- BLAU, P.M. (1964). *Exchange and Power in Social Life*. New York: John Wiley & Sons, Inc.
- CALLENS, M. (1995). *De 'Fertility and Family Survey' in Vlaanderen (Nego V, 1991). De gegevensverzameling*. Brussel: CBGS-document 1995, 4.
- CHAPMAN, D.W. (1983). The impact of substitution on survey estimates. Dans *Incomplete Data in Sample Surveys, Proceedings of the Symposium*, (Eds. W.G. Madow et I. Olkin). New York: Academic, 2, 45-61.
- CLIQUE, R.L. et CALLENS, M. (Ed.). (1993). *Gezinsvorming in Vlaanderen. Hoe en Wanneer?* Brussel: CBGS Monografie 1.
- COUPER, M.P., SINGER, E. et KULKA, R.A. (1997). Participation in the decennial census: Politics, privacy, pressures. *American Politics Quarterly*, 26, 59-80.
- GROVES, R.M. et COUPER, M. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.
- SAS INSTITUTE INC. (1999). *SAS/STAT® User's Guide*. Version 8, Cary, NC: SAS Institute Inc.
- SNIDERS, T.A.B., et BOSKERS, R.J. (1999). *Multilevel Analysis, an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications.
- VANNESTE, D. (1989). *Economische Typering Van de Belgische Gemeenten*. Leuven: Leuvense Geografische Papers.
- VEHOVAR, V. (1999). Field substitution and unit nonresponse. *Journal of Official Statistics*, 15, 2, 335-350.
- Doctoraatsbeurs 2002-2003 du FWO.

REMERCIEMENTS

Le présent article est le fruit de nos activités relatives à la conception et à l'organisation des enquêtes menées au Centre d'étude de la population et de la famille (CBGS), un institut scientifique de la Communauté flamande de Belgique. Nous sommes reconnaissants à l'Institut national de statistique de la Belgique, aux autorités locales et au Registre national de leur collaboration. Nous remercions aussi le rédacteur et deux examinateurs anonymes de leurs suggestions et commentaires constructifs. Enfin, l'étude a été financée par une bourse du programme Bijzonderere Doctoraatsbeurs 2002-2003 du FWO.

le pays. Ce genre d'hypothèse est nettement moins facile à vérifier, car elle nécessite des données sur plusieurs pays. À l'étape de l'analyse par régression logistique multi-niveaux, l'effet de tous les facteurs contextuels, sauf un, disparaît complètement. Seul l'effet de la variable de pourcentage d'immeubles à logements multiples présente une certaine résistance, quoique faible, au caractère aléatoire environnemental présent dans les modèles à ordonnée à l'origine aléatoire, résultat qui nous paraît fort curieux. La variation aléatoire environnementale au niveau municipal semble dominer en grande partie, même la dichotomie région urbaine-région rurale. Une explication éventuelle est que la variation au niveau communautaire est dominée par les effets d'intervieweur et non par les facteurs environnementaux.

La seule variable de niveau 2 ayant une importance (modérée) est celle des immeubles à logements multiples (en pourcentage), qui a été retenue dans le modèle 4. Le test du rapport des vraisemblances pour l'introduction de cette variable donne une différence de deux unités en ce qui concerne la somme des carrés des écarts à la moyenne. L'introduction d'une ou de plusieurs autres variables de niveau 2 donne des différences entre les résultats des tests qui concerne la somme des carrés des écarts. Nous considérons les modèles 3 et 4 comme étant les plus appropriés. La différence entre les sommes des carrés des écarts à la moyenne des modèles 3 et 2 est de huit unités, résultat qui est significatif. La variance du terme d'ordonnée à l'origine est modérée (0,21). L'introduction de covariables de deuxième niveau (y compris celle d'immeubles à logements multiples) ne modifie pour ainsi dire pas ce terme de variance. Par conséquent, nous pouvons déclarer que les attributs environnementaux, comme l'urbanité, ne jouent pas un rôle important dans l'explication de la coopération.

5. DISCUSSION

Le présent article décrit l'utilisation de données du Recensement de 1991 de niveaux individuel et municipal appariées aux données sur la variable de situation de réponse de l'Enquête sur la fécondité et la famille réalisée en Belgique pour analyser l'importance relative des corrélats de la prise de contact et de la coopération.

L'analyse est structurée conformément au cadre conceptuel de Groves-Couper. À l'étape de l'analyse bivariable, nous observons essentiellement le même type de corrélats que ceux prévus et effectivement observés lors d'une analyse américaine portant sur plusieurs enquêtes (G et C 1998). Une différence importante semble tenir à la nature de l'effet des indicateurs de SSB (par exemple, le niveau d'études) sur la coopération. Nous constatons une relation positive alors que l'étude américaine a abouti à l'observation d'une relation inverse. Nous pouvons imaginer deux explications différentes de ces résultats contradictoires. La première est fondée sur les effets de la conception de l'enquête, comme l'importance du sujet. L'EFF de la Belgique est peut-être atypique en ce sens qu'elle pourrait intéresser disproportionnellement la population la plus instruite à cause de son contenu particulier. La répétition de la présente analyse pour des enquêtes portant sur des sujets variés permettrait de tester facilement cette hypothèse. Une autre hypothèse est que les effets du niveau d'études sur la collaboration à l'enquête varient d'une société à l'autre. S'il en est ainsi, le défi consiste à découvrir pourquoi cette relation varie selon

niveau secondaire – deuxième cycle c. niveau supérieur), l'âge (20 à 29 ans, 30 à 39 ans) et l'échantillon de substitution. Nous avons choisi l'aire des pièces de séjour et le niveau d'études comme seuls indicateurs de l'hypothèse de l'échange social dans le modèle de régression logistique multiple décrit antérieurement. L'âge est la seule variable dont l'effet est significatif pour les résultats bivariables de l'hypothèse d'isolement social. Enfin, la substitution est introduite pour tenir compte des effets éventuels du travail sur le terrain. L'effet légèrement négatif de la substitution dans le modèle 2 pourrait indiquer que la substitution sur le terrain a un effet négatif sur la coopération. Cependant, cet effet disparaît entièrement si l'on introduit une ordonnée à l'origine aléatoire (modèles 3 et 4). Les effets des autres covariables de niveau individuel concordent avec les résultats de l'analyse bivariable et sont les mêmes pour les modèles 2 à 4. Les indicateurs de SSB, comme le niveau d'études et l'aire des pièces de séjour ont un effet positif et l'âge a un effet négatif sur la coopération. Ces effets confirment l'hypothèse de l'isolement social plutôt que celle de l'échange social.

Tableau 9

Résultats des modèles de régression logistique (multiniveaux) de la coopération

Résultats	Modèle 1 : aléatoire	Modèle 2 : régression aléatoire	Modèle 3 : aléatoire	Modèle 4 : régression aléatoire
	nul	logistique	aléatoire	aléatoire
	niveau 1	niveau 1	niveau 1	niveau 1
	et 2			
Ordonné à l'origine	1,41***	1,24***	1,30***	1,39***

Covariétés individuelles	-0,15*	-0,03	-0,02	(0,06)	(0,06)	(0,06)
Echantillon de substitution	(0,07)	(0,07)	(0,07)	0,24***	0,24***	0,24***
L'aire des pièces de séjour	(0,06)	(0,06)	(0,06)	0,47***	0,47***	0,47***
Niveau d'études	(0,08)	(0,08)	(0,08)	-0,23***	-0,23***	-0,23***
Âge	(0,06)	(0,06)	(0,06)	-0,006	-0,006	-0,006
Covariétés municipales	(0,004)	(0,004)	(0,004)	0,21	0,21	0,21
Var (ordonnée à l'origine)	0,21	0,21	0,21	6 664	6 596	6 594
Qualité de l'ajustement	6 664	6 664	6 664	6 594	6 594	6 594
Somme des carrés des écarts	6 664	6 664	6 664	6 594	6 594	6 594

Nota : Les erreurs-types sont présentées entre parenthèses. * $p < 0,05$, ** $p < 0,01$, *** $p < 0,001$, tests unilatéraux.

La variable de contrôle de substitution a un effet légèrement négatif sur la coopération ($\chi^2 = 4,24$, $p = 0,039$), le taux de coopération étant plus faible pour l'échantillon de substitution (77,3 %) que pour l'échantillon de base (79,5 %).

Le tableau 8 contient les résultats bivariés pour les différences de coopération selon l'environnement social. La densité de population a un effet curvilinéaire sur la coopération. Le fait de résider dans une grande région métropolitaine n'a aucun effet. Donc, les preuves données dans la littérature que le surpeuplement et le niveau élevé de stimulation sont négativement associées à la coopération sont discutables.

Tableau 8

Pourcentage de coopération selon l'attribut « environnemental »

Attribut environnemental	Pourcentage de coopération	χ^2	ddl	p
--------------------------	----------------------------	----------	-----	---

Situation d'urbanité		0,84	1	0,36
Ville	80,1			
Autre	78,7			
Densité de population		10,7	3	0,014
Quantile inférieur	80,0			
Deuxième quantile	79,9			
Troisième quantile	76,0			
Quantile supérieur	79,4			
Pourcentage d'immeubles à logements multiples		3,1	3	0,38
Quantile inférieur	80,1			
Deuxième quantile	79,2			
Troisième quantile	77,9			
Quantile supérieur	78,1			
Pourcentage de logements occupés par le propriétaire		12,3	3	0,0063
Quantile inférieur	79,7			
Deuxième quantile	76,2			
Troisième quantile	78,5			
Quantile supérieur	80,9			
Pourcentage de personnes d'autres nationalités		5,2	3	0,16
Quantile inférieur	77,7			
Deuxième quantile	77,6			
Troisième quantile	79,6			
Quantile supérieur	80,2			

L'effet des indicateurs de cohésion sociale n'est pas clair. Seule la variable de pourcentage de logements occupés par le propriétaire a un effet (curvilinéaire). Les variables de pourcentage de personnes d'autres nationalités et de pourcentage d'immeubles à logements multiples semblent n'avoir aucun effet. Enfin, nous présentons au tableau 9 une série de modèles de régression pour la coopération comparable à ceux de la section 4.1. Dans ces modèles, nous avons inclus quatre covariables de niveau individuel, à savoir l'aire des pièces de séjour (<84, >84 m²), le niveau d'études (jusqu'au

Pourcentage de coopération selon les indicateurs de l'« hypothèse d'isolement social »

Indicateurs d'isolement social	Pourcentage de coopération	χ^2	ddl	p
--------------------------------	----------------------------	----------	-----	---

Sexe		1,56	1	0,21
Masculin	78,1			
Féminin	79,3			
État civil		3,11	2	0,21
Non-marié	79,8			
Marié	78,6			
Divorcé	75,4			
Immeuble unifamilial		0,76	1	0,38
Non	78,9			
Oui	77,7			
Age		17,5	3	0,0006
20-24	80,8			
25-29	80,7			
30-34	78,3			
35-39	75,5			
Nombre d'enfants ^a		18,2	3	0,0004
0	77,9			
1	76,3			
2	81,7			
3+	84,9			
Présence de jeunes enfants ^a		12,3	1	0,0005
Non	77,8			
Oui	82,8			

^a Sous-échantillon de femmes uniquement ($n = 3\ 955$)

À la section sur l'hypothèse de l'échange social, nous dégageons certaines preuves que les personnes dont le SSE est faible coopèrent moins aux enquêtes que celles dont le SSE est élevé. Ce genre de relation positive entre le SSE et la coopération est prévu par l'hypothèse d'isolement social. Les indicateurs démographiques de la théorie de l'isolement social sont le sexe, l'état civil et l'âge (voir le tableau 7). Nous n'observons aucun effet pour le sexe, l'état civil (cependant, les personnes divorcées sont probablement moins coopératives) ni la vie dans un immeuble unifamilial. L'âge semble avoir un effet négatif sur la coopération. Pour les femmes uniquement, nous disposons aussi de données sur la présence d'enfants dans le ménage. Nous constatons que le nombre d'enfants a un effet positif sur le taux de coopération. L'âge des enfants est également important, la présence de jeunes enfants étant associée à un taux plus élevé de coopération.

Dans un modèle de régression logistique multiple des effets combinés des indicateurs d'échange social qui ont un effet bivariate marginal sur la coopération, seuls les effets du niveau d'études ($\chi^2 = 39,35$, $ddl = 4$, $p < 0,0001$) et l'aire des pièces de séjour ($\chi^2 = 13,4$, $ddl = 4$, $p = 0,0095$) demeurent significatifs.

ménage. Le résultat pour le niveau d'études est une autre indication d'une association positive entre la coopération et le SSE. Le fait de recevoir ou non un revenu de remplacement de l'Etat et d'être ou non propriétaire du logement n'ont pas d'effet sur le taux de coopération.

Tableau 5
Pourcentage de coopération selon les indicateurs de l'« hypothèse du coût de renonciation »

Indicateurs de coût de renonciation		cooperated		χ^2		ddl		p	
Inactif(ve) c. autre		77,0		0,41		1		0,52	
Autre		78,9							
Temps partiel c. temps plein ^b		82,3		10,04		1		0,001	
Temps plein		77,4							
Nombre d'heures travaillées ^b		80,1		15,3		3		0,0016	
<20		84,7							
21-35		77,6							
36-42		75,7							
>43		78,7		4,2		1		0,04	
Situation d'emploi ^b									
Employé		74,6							
Travailleur autonome									

^b Sous-échantillon de personnes actives uniquement ($n = 5\ 180$)

Tableau 6
Pourcentage de coopération selon les indicateurs de l'« hypothèse d'échange social »

Indicateurs d'échange		cooperation		χ^2		ddl		p	
Aire des pièces de séjour (m ²)		26,8		4		<0,0001			
< 65		74,8							
65 à 84		77,6							
85 à 104		78,6							
105 à 124		79,9							
> 125		83,1							
Nombre de salles de bain		74,2		7,9		2		0,02	
0									
1		78,6							
2		83,5							
Niveau de scolarité		76,6		46,7		4		<0,0001	
Primaire									
Secondaire, premier cycle		74,5							
Secondaire, deuxième cycle		78,7							
Supérieur, non universitaire		85,1							
Supérieur, universitaire		82,2							
Revenu de remplacement		78,7		0,3		1		0,58	
Non									
Oui		79,5							
Logement occupé par le propriétaire		77,4		3,4		1		0,06	
Non									
Oui		79,4							

tandis que l'effet de la situation d'activité n'est pas significatif. L'effet de la substitution sur le terrain est négatif. Nous notons aussi une réduction (assez faible) de la valeur du coefficient de régression pour l'habitation d'un immeuble unifamilial et la substitution dans les modèles multivariés 3 et 4. Ces deux modèles ont une composante de variance pour l'ordonnée à l'origine. Pour tester l'hypothèse nulle selon laquelle la variance à l'ordonnée à l'origine est nulle, nous utilisons le test du rapport des vraisemblances et comparons le modèle logit classique (modèle 2) au modèle à ordonnée à l'origine aléatoire (modèle 3). La différence entre les sommes des carrés des écarts des deux modèles est importante (52). Par conséquent, il pourrait exister dans l'ordonnée à l'origine une certaine variance que l'on devrait expliquer au moyen de covariables de niveau municipal. En introduisant les caractéristiques de la municipalité une à la fois, nous pouvons tester quels effets sont significatifs en calculant les différences entre les sommes des carrés des écarts du modèle 4 et du modèle 3. La seule différence importante est celle observée pour la variable « immeuble à logements multiples » (sept unités de différence). Aucune différence entre les sommes des carrés des écarts n'est observée pour l'introduction des autres variables de niveau 2 (situation d'urbanité, pourcentage de logements occupés par le propriétaire, densité de population et pourcentage de personnes d'autres nationalités).

Nous considérons que les modèles 3 et 4 sont les meilleurs. Selon ces modèles multivariés, le taux de non-prise de contact varie considérablement selon la municipalité. Cependant, les covariables de niveau municipal considérées dans notre étude ne permettent d'expliquer qu'une faible part de cette variation.

4.2 Coopération

Au tableau 5, nous présentons les résultats bivariés pour les indicateurs de l'hypothèse du coût de renonciation. Le fait d'être inactif ou non ne semble pas avoir d'effet sur le taux de coopération. Cependant, si nous utilisons des indicateurs de temps de loisir, comme le travail à temps partiel par opposition au travail à temps plein ou le nombre d'heures travaillées par semaine, les résultats bivariés confirment la relation négative prévue. En outre, le taux de coopération est plus faible pour les travailleurs autonomes que pour les employés.

Les prévisions fondées sur la théorie de l'échange social ne s'observent pas dans les résultats présentés au tableau 6. Les indicateurs de SSE, comme l'aire des pièces de séjour et le nombre de salles de bain, sont associés non pas négativement, mais positivement, à la coopération. Naturellement, ces mesures ne sont pas parfaites, parce qu'il nous est impossible de tenir compte de la taille du

base (2,6 %). Aucune différence significative n'est observée pour les sous-échantillons d'hommes et de

fermes.

Dans un modèle de régression logistique multiple des effets combinés des indicateurs au niveau individuel ayant

un effet bivariable marginal sur la possibilité de prise de contact, seul les attributs immeuble unifamilial ($\chi^2 = 35,75$, $p \leq 0,0001$), téléphone ($\chi^2 = 52,63$, $p \leq 0,0001$) et substitution

($\chi^2 = 28,59$, $p \leq 0,0001$) restent significatifs.

Au tableau 3, nous présentons les taux de non-prise de contact pour divers attributs environnementaux. Le taux de non-prise de contact est plus élevé pour les villes (6,6 %) que pour les régions non urbaines (3,1 %). Le pourcentage de personnes avec lesquelles il n'est jamais pris contact est plus élevé pour les zones à forte densité de population (5,4 %) que pour celles à faible densité (1,7 %). L'existence d'immeubles à logements multiples et la présence de personnes d'autres nationalités ont tendance à faire augmenter le taux de non-prise de contact. Enfin, il existe une association négative entre le pourcentage de logements occupés par le propriétaire et le taux de non-prise de contact.

Tableau 3
Pourcentage de personnes qui n'ont jamais été rejointes selon l'attribut « environnemental »

Attribut environnemental	Pourcentage de personnes jamais rejointes	χ^2	df	p
Situation d'urbanité	24,0	1	<0,0001	
Ville	6,6			
Autre	3,1			
Densité de population	1,7	34,4	3	<0,0001
Quartile inférieur	3,2			
Deuxième quartile	3,8			
Troisième quartile	5,4			
Quartile supérieur	5,4			
Logements multiples	50,4	3	<0,0001	
Quartile inférieur	2,0			
Deuxième quartile	2,2			
Troisième quartile	4,0			
Quartile supérieur	5,9			
Pourcentage de personnes d'autres nationalités	23,1	3	<0,0001	
Quartile inférieur	2,5			
Deuxième quartile	2,3			
Troisième quartile	4,3			
Quartile supérieur	4,8			
Pourcentage de logements occupés par le propriétaire	64,4	3	<0,0001	
Quartile inférieur	6,4			
Deuxième quartile	3,6			
Troisième quartile	1,6			
Quartile supérieur	2,7			

Tableau 4
Résultats des modèles de régression logistique (multiniveaux) de la possibilité de prise de contact

Résultats	Modèle 1 : aléatoire nul	Modèle 2 : régression aléatoire	Modèle 3 : régression aléatoire	Modèle 4 : régression aléatoire
Ordonné à l'origine	4,01***	3,08***	3,68***	4,15***
Caractéristiques individuelles	(0,16)	(0,73)	(0,77)	(0,79)
Immeuble unifamilial	1,16***	1,02***	0,92***	
Caractéristiques municipales	(0,15)	(0,17)	(0,17)	
Téléphone	1,19***	1,25***	1,26***	
Inactif(ve) c. autre	(0,16)	(0,17)	(0,17)	
Echantillon de substitution	(0,72)	(0,75)	(0,74)	
Caractéristiques municipales	(0,14)	(0,15)	(0,15)	
Immeubles à logements multiples	-0,78***	-0,64***	-0,62***	
Var (ordonnée à l'origine)	1,03	0,82	0,79	
Qualité de l'ajustement				
Somme des carrés	1 720	1 658	1 606	1 599
des écarts				

Nota: Les erreurs-types sont présentées entre parenthèses. * $p < 0,05$, ** $p < 0,01$, *** $p < 0,001$, tests unilatéraux.

Les effets des covariables au niveau de la personne introduites dans les modèles 2, 3 et 4 concordent avec les observations de l'analyse bivariable. Le fait de vivre dans un immeuble unifamilial et l'existence d'un téléphone ont une influence positive sur la possibilité de prise de contact,

un enfant (4 %). Le taux de non-prise de contact est également un peu plus faible pour les personnes qui travaillent à domicile (1,5 %) et pour celles qui travaillent à leur propre compte (1,9 %) que pour celles qui travaillent ailleurs (3,6 %) ou qui sont employées (3,6 %), respectivement. L'âge, le nombre d'heures travaillées par semaine, le travail à temps partiel par opposition au travail à temps plein et le fait d'avoir ou non un deuxième emploi n'ont aucun effet significatif sur la possibilité de prise de contact.

Tableau 1

Pourcentage de personnes qui n'ont jamais été rejointes selon le type d'« obstacle physique »

Type d'obstacle physique	Pourcentage de personnes jamais rejointes	χ^2	<i>ddl</i>	<i>p</i>
Immeuble unifamilial	Non	97,6	1	<0,0001
	Oui	8,1		
Grand immeuble à logements multiples (>10)	Non	38,4	1	<0,0001
	Oui	3,1		
Téléphone	Non	88,9	1	<0,0001
	Oui	2,7		

Tableau 2

Pourcentage de personnes qui n'ont jamais été rejoints selon la « raison de la présence au domicile »

statistiques	Personnes seules	Personnes avec un conjoint	Personnes avec un ou deux conjoint(e)s
État civil	19,4	2	<0,0001
Non marié(e)	4,4		
Marié(e)	2,9		
Divorcé(e)	6,9		
Inactif(ve) c. autre	4,0	1	0,04
Inactif(ve)	0,9		
Autre	3,5		
Nombre d'enfants ^a	14,5	3	0,0023
0	4,3		
1	4,0		
2	2,6		
3+	0,9		
Lieu d'emploi ^b	4,6	1	0,03
À domicile	1,5		
Villeurs	3,6		
Type d'emploi ^b	4,0	1	0,05
Employé(e)	3,6		
À son propre compte	1,9		

Sous-échantillon de femmes uniquement ($n=4\ 098$)

En incluant également s variables au niveau communautaire, nous obtenons un modèle à ordonnée à l'origine comprenant des covariables de niveau 1 et de niveau 2 :

des variables explicatives. Si nous laissons tomber les effets aléatoires u_{0j} , nous obtenons un modèle de régression logistique standard :

$$f_0 n + f_1 y x^y \lambda \sum_{\lambda}^{l=y} + {}^0 \lambda = (f_1 d) \log 10$$

où $\gamma_1, \dots, \gamma_r$ sont les paramètres de pente mesurant l'effet des variables explicatives.

$$\cdot^{f_1} x^y \lambda \sum_{j=1}^{l=y} + {}^0 \lambda = ({}^{f_1} d) \log$$

$$f_0 \mathbf{n} + f_{1x} x^x \lambda \sum_{\lambda=1}^{\lambda=x} + f_{1y} x^y \lambda \sum_{\lambda=1}^{\lambda=y} + 0 \lambda = (f_1 d) \mathbf{m} \otimes \mathbf{I}$$

Nous utilisons SAS Proc Nlmixed (SAS Institute, 1999) pour estimer les paramètres. Dans SAS Proc Nlmixed, une version adaptée de la quadrature de Gauss-Hermite (intégration numérique) est utilisée pour résoudre le problème d'estimation du maximum de vraisemblance. Pour vérifier si un paramètre particulier est nul, on utilise un test du χ^2 du rapport des vraisemblances.

4. RESULTS

4.1 Possibilité de prise de contact

Le tableau 1 donne les résultats bivariés pour le test du χ^2 du sous-échantage de personnes qui ont des difficultés à accéder au logement physique. Un corrélat important est le fait que le logement soit ou non un immeuble unifamilial, le taux de non-prise de contact étant nettement plus élevé pour les unités de logement dont la structure n'est pas unifamiliale (8,1 %) que pour les autres (2,4 %). En outre, le taux de non-prise de contact a tendance à être plus élevé pour les personnes échappées vivant dans de grands immeubles à logements multiples (11 %) que pour celles ne vivant pas dans ce genre d'immeuble (3,1 %). Un autre corrélat important est l'existence d'un téléphone, aucun contact n'ayant été établi avec 9,7 % de personnes non abonnées au téléphone.

Le tableau 2 donne les résultats bivariés pour la possibilité de prendre contact selon la « raison » d'être présent au domicile ». Une proportion relativement plus élevée de personnes échantillonnées non mariées (4,4 %) et divorcées (6,9 %) que de personnes mariées (2,9 %) ne sont jamais rejointes. Le taux de non-prise de contact est beaucoup plus faible pour les personnes inactives (0,9 %) que pour les autres (3,5 %). Le fait d'avoir au moins trois enfants donne un taux de non-prise de contact faible (0,9 %), comparativement au fait d'avoir deux enfants (2,6 %) ou au moins un enfant (4,2 %).

de substitution) et le sexe (le fait qu'une personne échantillonnée provienne de l'échantillon de femmes ou de l'échantillon d'hommes).

Au niveau municipal ($n=123$), nous utilisons cinq variables, à savoir la densité de population (nombre d'habitants par kilomètre carré pour le lieu de résidence de la personne échantillonnée), la situation urbaine (les villes d'Anvers et de Gand c. les autres municipalités), le pourcentage d'immuebles à logements multiples (par quartile : <7,13; 7,13 à 15,14; 15,14 à 27; >27), le pourcentage de logements occupés par le propriétaire (par quartile : <64,5; 64,5 à 71; 71 à 77,7; >77,7) et le pourcentage de personnes appartenant à une race minoritaire (par quartile : <0,90; 0,90 à 2,22; 2,22 à 5,29; >5,29).

3.1.4 Coopération et ses déterminants

Nous souhaitons déterminer la probabilité de jamais obtenir une coopération (code 1 = coopération et code 0 = non-coopération), conditionnellement à une prise de contact, et non s'il a été facile ou difficile d'obtenir la coopération de la personne échantillonnée. Pour 1 399 des 6 606 personnes échantillonnées avec lesquelles il y a eu prise de contact (21,18 %), toute tentative en vue d'obtenir leur coopération a échoué.

De nouveau, les données que nous utilisons sont mesurées à deux niveaux, à savoir ceux de l'individu et de la municipalité. Au niveau de la personne échantillonnée, nous disposons d'indicateurs pour l'hypothèse du coût de renonciation, l'hypothèse d'échange social et l'hypothèse d'isolement social. Nous utilisons la substitution comme variable de contrôle.

Pour l'hypothèse du coût de renonciation, les indicateurs sont la situation d'activité (inactif(ve) c. autre), le travail à temps partiel c. le travail à temps plein, le nombre d'heures travaillées par semaine (<21, 21 à 35, 36 à 42, >42 heures) et le type d'emploi (employé(e) c. à son propre compte).

Nos indicateurs de statut socioéconomique sont l'aire des pièces de séjour (en mètres carrés) (<65; 65 à 84; 85 à 104; 105 à 124; >125), le nombre de salles de bain (0, 1 et 2+) et le niveau d'études (primaire, secondaire – premier cycle, secondaire – deuxième cycle, supérieur – non universitaire, supérieur – universitaire). Les autres indicateurs situent l'hypothèse de l'échange social sont le fait de recevoir ou non un revenu de remplacement de l'Etat et le fait d'être ou non propriétaire du logement occupé.

Pour l'hypothèse de l'isolement social, les indicateurs sont le sexe, l'état civil (non marié(e), marié(e) ou divorcé(e)), l'âge (20 à 24 ans, 25 à 29 ans, 30 à 34 ans et 35 à 39 ans), le logement dans un immeuble unifamilial et pour les femmes uniquement, le nombre d'enfants (0, 1, 2 et 3+) et la présence d'enfants de moins de cinq ans. Enfin, la substitution est incluse comme variable de contrôle.

3.2.2 Régression logistique multivariée

Dans une deuxième série d'analyses, nous utilisons la régression logistique multivariée pour estimer simultanément l'effet de divers déterminants (Snijders et Boskers 1999). Nous optons pour une méthode multivariée, parce que nos données sont emboîtées hiérarchiquement, les individus correspondant au niveau inférieur (niveau 1) et les municipalités, au niveau supérieur (niveau 2). Soit p_{ij} la probabilité qu'un individu i appartenant à la municipalité j soit rejoint (ou coopère). Nous considérons quatre modèles distincts pour expliquer cette probabilité, à savoir le modèle aléatoire nul, deux versions du modèle à ordonnée à l'origine aléatoire et le modèle de régression logistique standard.

Le modèle vide ou non conditionnel ne tient pas compte des variables explicatives. Nous spécifions le modèle de la façon que les probabilités transformées $\text{logit } p_{ij}$ suivent une loi normale :

$$\text{logit}(p_{ij}) = 1 / (1 + \exp(D_{ij})) = \gamma_0 + n_{0j}$$

où γ_0 est la moyenne de population et n_{0j} , l'écart aléatoire par rapport à cette moyenne pour le groupe j . Nous supposons que les écarts n_{0j} sont des variables aléatoires indépendantes et normalement distribuées de moyenne nulle et de variance τ_0^2 . Quand il existe r variables au niveau individuel qui pourraient expliquer les résultats observés, elles sont intégrées sous forme de fonction linéaire dans le modèle à ordonnée à l'origine aléatoire :

3.1.3 Possibilité de prise de contact et ses déterminants

Idealement, pour étudier le processus de la possibilité de prise de contact, il faut posséder des données sur les résultats de toutes les tentatives successives de prise de contact avec les personnes échantillonnées. Pour la présente étude, nous ne disposons malheureusement pas de renseignements aussi détaillés; nous ne connaissons que le résultat final de chaque demande de participation à l'enquête. Par conséquent, nous pouvons étudier la probabilité de jamais prendre contact avec la personne échantillonnée (code 1 = prise de contact et code 0 = non-prise de contact), mais non s'il a été facile ou difficile d'établir le contact. Nous considérons les personnes échantillonnées que l'on sait ne pas (ou ne plus) résider à l'adresse échantillonnée comme ayant été rejointes. Pour 241 des 6 847 unités échantillonnées (3,52 %), toutes les tentatives de prise de contact ont échoué.

Les données que nous utilisons sont produites à deux niveaux, c'est-à-dire les niveaux individuel ($n=6\,847$) et municipal ($n=123$). Au niveau de la personne échantillonnée, nous considérons trois types de variables, à savoir les obstacles physiques à la prise de contact avec la personne échantillonnée, les raisons pour lesquelles les personnes échantillonnées sont présentes à leur domicile et les variables de contrôle.

Comme nous ne disposons d'aucune observation directe des interveneurs sur les obstacles physiques, nous devons nous appuyer uniquement sur les indicateurs d'obstacle physique qui existent dans les données du recensement. Les trois variables utilisées sont le fait que l'unité de logement soit ou non un immeuble unifamilial, le fait que l'immeuble comprenait l'unité de logement soit grand (plus de 10 unités) ou non, et le fait que la personne échantillonnée soit abonnée ou non au téléphone.

Dans la présente étude, les déterminants du profil de présence au domicile sont l'état civil (non-marié(e), marié(e) et divorcé(e)), l'âge (20 à 24 ans, 25 à 29 ans, 30 à 34 ans, et 35 à 39 ans) et la situation d'activité (inactif(ve) c. autre). Pour les femmes uniquement, nous considérons aussi le nombre d'enfants (0, 1, 2 et 3+). Pour les personnes faisant partie de la population active, nous avons aussi des renseignements détaillés sur le travail à temps partiel vs. travail à temps plein, le nombre d'heures travaillées par semaine (<21, 21 à 35, 36 à 42, >42 heures), le type d'emploi (employé(e) c. à son propre compte), l'existence d'un deuxième emploi ou non, et le travail à domicile ou non.

Nous utilisons aussi deux variables de contrôle, à savoir la substitution (le fait qu'une personne échantillonnée provienne de l'échantillon cible de base ou de l'échantillon

(Chapman 1983; Vehovar 1999). L'échantillon final, c'est-à-dire y compris l'opération de substitution, comprend 4 776 personnes (2 897 femmes et 1 879 hommes). Dans la présente étude, nous utilisons les cas de réponse et de non-réponse de l'échantillon cible initial, ainsi que de l'opération de substitution sur le terrain ($N=6\,847$).

Aussi bien chez les hommes que chez les femmes, dans sept cas sur dix, la non-réponse peut être attribuée à un refus de participer à l'enquête. Dans deux cas sur dix, elle est due au fait qu'il a été impossible de prendre contact avec la personne sélectionnée et dans un cas sur dix, l'interview n'a pas été possible à cause d'une maladie, de difficultés linguistiques ou d'autres raisons.

3.1.2 Appariement des données au niveau de la personne du Recensement de 1991 (1991)

Notre source principale d'information sur les cas de réponse et de non-réponse est le Recensement de 1991.

Pour nous efforcer de concilier la protection des renseignements personnels et les intérêts scientifiques, nous avons utilisé une technique simple pour appairer les données du recensement au niveau de la personne aux données d'enquête anonymes. Nous avons fourni à l'Institut national de statistique (INS) un ensemble de données contenant uniquement le numéro d'identification national et la situation de réponse pour chaque cas répondant et non répondant. Suite à l'opération d'appariement réalisée par l'INS, nous avons reçu une sélection de données du Recensement de 1991 enrichies au moyen de données sur deux variables d'enquête uniquement, à savoir la variable de situation de réponse et un indicateur précisant si la personne fait partie de l'échantillon de base ou de l'échantillon de substitution.

Les données de niveau individuel du Recensement de 1991 dont nous disposons sont celles du questionnaire individuel et du questionnaire sur l'unité de logement. Le questionnaire individuel contient des renseignements sur le lieu de résidence, la situation d'activité, le premier mariage, l'année de naissance des enfants, ainsi que les activités scolaires et professionnelles. Le questionnaire sur l'unité de logement comprend des renseignements sur l'unité de logement du ménage, dont le type d'unité de logement, le nombre d'unités de logement dans l'immeuble, la propriété, la date de construction, le nombre de pièces et le nombre de mètres carrés correspondant, la présence d'un téléphone et des indicateurs de confort, comme le nombre de salles de bain.

rurales a une tendance à être plus élevée que celui des habitants des villes. Cependant, le mécanisme responsable de cet effet d'urbanité reste à préciser. Il pourrait être dû à la plus forte densité de population, le taux plus élevé de criminalité et le niveau plus élevé de désorganisation sociale associés à la vie dans les régions urbaines. On suppose que la densité de population réduit la coopération à cause de l'expérience de surpeuplement. La crainte de la criminalité peut être à l'origine du refus de fournir des renseignements à des étrangers. Enfin, la vie urbaine est associée à la désorganisation sociale, caractérisée par l'affaiblissement des réseaux locaux de parenté et d'amitié et la réduction de la participation aux affaires locales.

3. DONNÉES ET MÉTHODE

3.1 Données

La présente étude porte à la fois sur des données agrégées et des microdonnées provenant du Recensement de la population de la Belgique de 1991 couplées à la variable de situation de réponse pour les répondants et les non-répondants à l'Enquête sur la fécondité et la famille en Belgique (EFF-Belgique 1991) réalisée peu après les opérations du recensement.

3.1.1 L'EFF de 1991

L'Enquête sur la fécondité et la famille en Belgique a été organisée par le Centre d'étude de la population et de la famille (CBGS), un institut scientifique créé par le gouvernement de la Communauté flamande. La collecte des données a eu lieu d'avril à octobre 1991, période qui est très proche de la date du recensement décennal, à savoir le 1^{er} avril de la même année. Le projet de l'EFF met principalement l'accent sur le comportement de procréation, qu'il convient toutefois d'examiner dans le contexte général des antécédents d'union consensuelle et d'antécédents familiaux, et de l'interaction entre l'activité économique et la procréation (Clquet et Callens 1993; Callens 1995). La population cible comprend les hommes et les femmes de nationalité belge, nés entre 1951 et 1970 et dont la résidence principale se situe dans la région flamande de Belgique.

Un plan d'échantillonnage en grappes à deux degrés a été utilisé pour échantillonner séparément les hommes et les femmes. Lors d'une première étape, les municipalités ont été sélectionnées à partir de diverses strates socio-économiques (Vaneste 1989). Puis, dans chaque municipalité sélectionnée, des individus ont été échantillonnés au hasard. De cette façon, 2 975 femmes et 1 989 hommes ont été sélectionnés pour participer à l'enquête. Une méthode sur le terrain a été appliquée pour compenser

publique qu'ils reçoivent. Les groupes à SSE plus élevés, quant à eux, se sentent nettement moins redevables de quoi que ce soit. Dans cette perspective, la relation entre le statut socioéconomique et la propension à coopérer est négative. D'autre part, on peut émettre l'hypothèse d'une relation curvilinearé entre le SSE et la coopération. Selon cette hypothèse, les groupes à SSE faible considèrent qu'ils sont systématiquement défavorisés comparativement aux personnes plus fortunées, tandis que les groupes à SSE élevés ont le sentiment que leur temps et leur argent sont sans cesse sollicités, mais qu'ils reçoivent peu en retour. Sous cette hypothèse, les groupes ayant le SSE le plus élevé et le SSE le plus faible se sentent l'un et l'autre relativement défavorisés dans leur relation avec les grandes institutions sociales et ont tendance à refuser de coopérer aux enquêtes.

2.2.3 Isolement social

L'hypothèse de l'isolement social est étroitement liée à l'hypothèse de l'échange social. Les personnes socialement isolées sont déphasées par rapport à la culture dominante d'une société et ont tendance à se comporter selon des normes sous-culturelles ou à rejeter explicitement celles de la culture dominante. Elles sont considérées comme étant moins susceptibles que les autres de participer à diverses activités sociales et politiques, y compris répondre à des enquêtes (Copper, Singer et Kulika 1997). En ce qui concerne le SSE, la théorie de l'isolement social implique une relation positive entre le SSE et la coopération, les groupes dont le SSE est faible éprouvant du ressentiment à l'égard de dépendre de l'État et ceux dont le SSE est élevé manifestant un sens plus prononcé de devoir civique. Cette relation positive entre le SSE et l'isolement social est l'opposé de celle prévue par la théorie

2.2.4 Urbanité

Au niveau communautaire, nous émettons l'hypothèse que les facteurs contextuels tels que l'urbanité, la densité de population, le taux de criminalité et le manque de cohésion sociale exercent une influence sur la participation à une enquête. Le taux de coopération des résidents des régions

Groves et Couper (G et C 1998) ont proposé un cadre théorique multinitiveaux séduisant pour étudier la possibilité de prise de contact et la coopération.

2.1 Possibilité de prise de contact

Chronologiquement, le processus de prise de contact avec une personne échantillonnée est celui qui survient en premier lieu. Certaines de ces personnes ne sont jamais

rejointes par les intervieweurs et, donc, ne prennent jamais de décision quant à leur participation à l'enquête. Comparativement au processus de coopération, le processus de prise de contact avec une personne échantillonnée est assez simple.

G et C (1998) considèrent que la possibilité de prise de contact, ou « contactabilité », dépend de trois facteurs, à savoir 1) l'existence ou non d'obstacles physiques qui empêchent l'intervieweur d'entrer en contact avec la personne échantillonnée, 2) la présence ou non de la personne échantillonnée à son domicile et 3) le moment et le nombre de fois que l'intervieweur essaye de prendre contact avec la personne échantillonnée. Le nombre et le moment des appels faits par l'intervieweur et le profil d'accessibilité de la personne échantillonnée à son domicile sont les causes proximales de la possibilité de prise de contact. Le profil d'accessibilité de la personne échantillonnée à son domicile dépend de l'existence d'obstacles physiques (par exemple présence d'un téléphone), d'attributs sociodémographiques (par exemple, le temps de déplacement pour se rendre au travail et en revenir) et d'attributs socioenvironnementaux (par exemple, criminalité). En outre, certaines caractéristiques du plan d'enquête, comme la durée de la période de collecte des données et la charge de travail de l'intervieweur, peuvent influencer sur les taux de prise de contact.

2.2 Coopération

À l'étape qui suit la prise de contact, la question centrale est de savoir pourquoi les personnes échantillonnées acceptent ou non de coopérer avec l'intervieweur. Dans le modèle élaboré par Groves-Couper pour étudier la coopération, les causes proximales de la décision de coopérer ou de refuser de participer se situent au niveau du chef du ménage et de l'interaction de celui-ci avec l'intervieweur. Une autre composante du cadre théorique de G et C (1998) est l'ensemble de caractéristiques de conception de l'enquête, comme les moyens de collecte des données, l'annonce préalable de la demande de participation à l'enquête, l'importance du sujet, etc.

G et C (1998) tiennent également compte de deux facteurs ne relevant pas du contrôle du concepteur de l'enquête, à savoir l'influence exercée par la personne échantillonnée et par le contexte socioenvironnemental. Ils

2.2.2 Échange social

La théorie de l'échange social considère la valeur perçue de l'équité des associations de longue durée entre personnes ou entre une personne et des institutions sociales (Blau 1964). Au cœur de toutes les conceptualisations de l'échange social, on trouve la notion selon laquelle, contrairement aux échanges économiques, tous les produits sociaux font partie d'un système de comptabilité intuitif dans lequel les dettes (par exemple, les obligations) et les crédits (par exemple les attentes) sont pris en compte (G et C 1998). La perspective de l'échange social peut être appliquée dans tous les cas où existe une relation courante entre l'organisme d'enquête et la personne échantillonnée (par exemple, enquête gouvernementale).

Les personnes qui reçoivent peu de services gouvernementaux pourraient, si l'on considère l'effet cumulatif de contacts multiples avec le gouvernement, se sentir moins obligées de coopérer. Puisque les services gouvernementaux sont prodigués de façon inégale dans les diverses strates socioéconomiques, les indicateurs du statut socio-économique (SSE) devraient refléter l'influence des échanges sur la participation aux enquêtes. Toutefois, un problème important de la théorie de l'échange social est qu'on peut en déduire deux hypothèses alternatives du lien entre le SSE et la coopération (G et C 1998). D'une part, on peut soutenir que les groupes dont le SSE est faible ont sans doute la plus grande dette envers l'Etat en raison de l'aide

La notion de coût de renonciation sous-entend que les personnes échantillonnées évaluent le coût de renonciation associé au fait d'accepter de consacrer de leur temps à répondre à une enquête. Un élément important de la théorie du coût de renonciation est le temps de loisir dont dispose la personne échantillonnée pour participer à l'enquête. Les personnes dont le temps de loisir est limité sont moins susceptibles de se sentir libre de participer à une enquête. Certains indicateurs indirects de la quantité de temps de loisir sont l'inverse du nombre d'adultes dans le ménage et la (quantité de) participation au marché du travail. Naturellement, il existe d'autres obligations que celles associées à l'emploi, comme les engagements à l'égard d'amis et de parents, qui sont également susceptibles d'accroître le coût de renonciation lié à la participation à une enquête.

2.2.1 Coût de renonciation

isolement social.

Importants sont le coût de renonciation, l'échange social et sociopsychologiques. À cet égard, des constructus théoriques donnent une mesure indirecte de constructus essentiellement influence causale directe sur la coopération, mais qu'elles considèrent non pas que ces variables exercent une

Prise de contact et coopération dans l'Enquête belge sur la fécondité et la famille

MARC CALLENS et CHRISTOPHE CROUX

RÉSUMÉ

Des données recueillies lors de l'Enquête sur la fécondité et la famille réalisée en Belgique sont combinées à des données sur les répondants et les non-répondants aux niveaux individuel et municipal provenant du Recensement de la population de la Belgique de 1991 pour estimer des modèles de régression logistique multivariés de la possibilité de prise de contact et de la propension à la coopération. Une sélection de caractéristiques indirectes, dont aucune ne peut être contrôlée directement par les chercheurs, sont introduites comme covariables. Contrairement aux études antérieures, nous constatons que le statut socioéconomique est positivement associé à la coopération. Un autre résultat inattendu est l'absence de tout effet important des corrélats environnementaux, tels que l'urbanité.

MOTS CLÉS : Non-réponse; analyse multivariés; enquête sur la fécondité et la famille.

1. INTRODUCTION

L'objectif du présent article est d'évaluer empiriquement l'importance relative des corrélats des taux de prise de contact et de coopération dans l'Enquête sur la fécondité et la famille en Belgique (EFF Belgique 1991).

Le cadre conceptuel et théorique de non-réponse utilise pour l'étude a été proposé par Groves et Couper (G et C 1998). Selon ces auteurs, la non-réponse due à l'absence de prise de contact est directement influencée par les caractéristiques de conception de l'enquête, comme le nombre d'appels et le moment de ceux-ci. Conditionnellement à ces caractéristiques de conception de l'enquête, d'autres caractéristiques importantes, comme les obstacles physiques que présentent les unités de logement et les profits d'accessibilité des répondants prospectés au domicile, qui sont les uns et les autres mesurés indirectement par divers attributs socioenvironnementaux et sociodémographiques, jouent aussi un rôle important. La décision de collaborer ou de refuser de participer à l'enquête est considérée avant tout comme une fonction directe d'un processus social de communication dynamique entre l'intervieweur et la personne interviewée. Les caractéristiques de conception de l'enquête, de l'interview principal, de la personne échantillonnée et de l'environnement social sont considérées comme ayant une influence indirecte sur les taux de coopération.

Nous utilisons des données de niveau individuel ainsi que municipal provenant du Recensement de 1991, appartenant à la variable de résultat du travail sur le terrain,

c'est-à-dire répondant ou non-répondant, de l'EFF belge de 1991. Dans cette enquête, l'unité d'échantillonnage est l'individu. Il s'agit d'une enquête avec interview sur place et le taux de refus de participation, modéré (22 %). Nous considérons que nos données sont emboîtées hiérarchiquement, les unités d'échantillonnage correspondant au niveau le plus faible et les municipales, au niveau le plus élevé. Nous incluons des covariables au deux niveaux et nous estimons des modèles de régression logistique multivariés pour la possibilité de prise de contact et la propension à la coopération. Les covariables sont une série de caractéristiques indirectes sur aucune desquelles les chercheurs n'exercent un contrôle direct. Certains résultats sont curieux, à savoir 1) les indicateurs de statut socioéconomiques, comme le niveau d'études, sont positivement corrélés à la coopération et 2) les facteurs environnementaux, y compris l'urbanité, ne sont pas corrélés à la non-réponse. Ces résultats contredisent ceux d'une étude réalisée antérieurement aux États-Unis.

2. UNE THÉORIE DE LA POSSIBILITÉ DE PRISE DE CONTACT ET DE LA COOPÉRATION

Le processus de réalisation d'une interview comprend deux éléments importants, à savoir le processus de prise de contact avec une personne échantillonnée et, sous réserve de cette prise de contact, le processus de coopération de la personne à qui l'on demande de participer à l'enquête.

- STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.
- STATISTIQUE CANADA (1998). Permanent layoffs, quits and hirings in the Canadian Economy: 1978-1995. Catalogue # 71-539-XIB.
- STATISTIQUE CANADA (1998). Enquête sur la dynamique du travail et du revenu - un aperçu de l'enquête. Catalogue # 75F0011XPB, <http://www.statcan.ca/francais/freepub/75F0011XIF/free.htm>.
- WANG, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86, 130-143.
- STATISTIQUE CANADA (2002). Guide de l'Enquête sur la population active. Catalogue # 71-543-GIF, <http://www.statcan.ca/francais/IPS/Data/71-543-GIF.htm>.
- STATISTIQUE CANADA (2001). Le modèle de microsimulation LifePaths : Une vue d'ensemble. <http://statcan.ca/francais/psd/LifePaths.htm>.

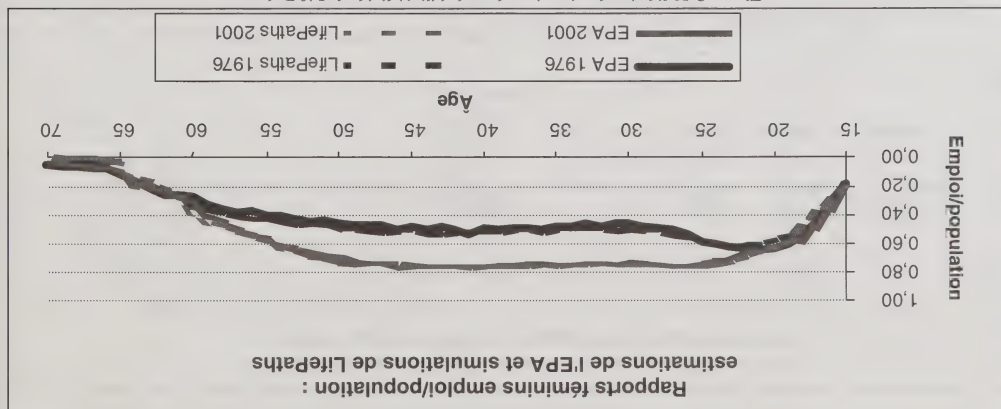


Figure 8. Validation des équations de probabilité à l'aide de LifePaths

5. CONCLUSIONS

Nous avons démontré que les données de l'EPA organisées en relevés fragmentaires sur les six mois que passent la plupart des répondants dans l'échantillon forment un important fonds de microdonnées longitudinales. L'échantillon et l'étendue du contenu sont suffisants pour une grande analyse de la dynamique du marché du travail et peut-être aussi d'aspects démographiques comme la fécondité. Ajoutons qu'il s'agit de données mensuelles sur plus d'un quart de siècle et que l'analyse aura donc une base temporelle ininterrompue qui est sans égale au Canada.

Dans notre application principale (passages d'emploi), d'autres résultats (que nous ne présentons pas) semblent confirmer l'incidence d'un éventail de variables explicatives sur les probabilités individuelles de passage, qu'il s'agisse de l'âge, de l'occupation d'emploi (ou de la durée d'absence (surtout dans le cas des femmes), de la province de résidence, de la saisonnalité ou des cycles économiques. Il reste que ces travaux en sont encore aux balbutiements et que, à ce stade, notre mode d'inférence est peu formel. Dans de futurs travaux, nous devrons étendre et affiner nos modèles et donner des bases plus rigoureuses à leur évaluation.

BIBLIOGRAPHIE

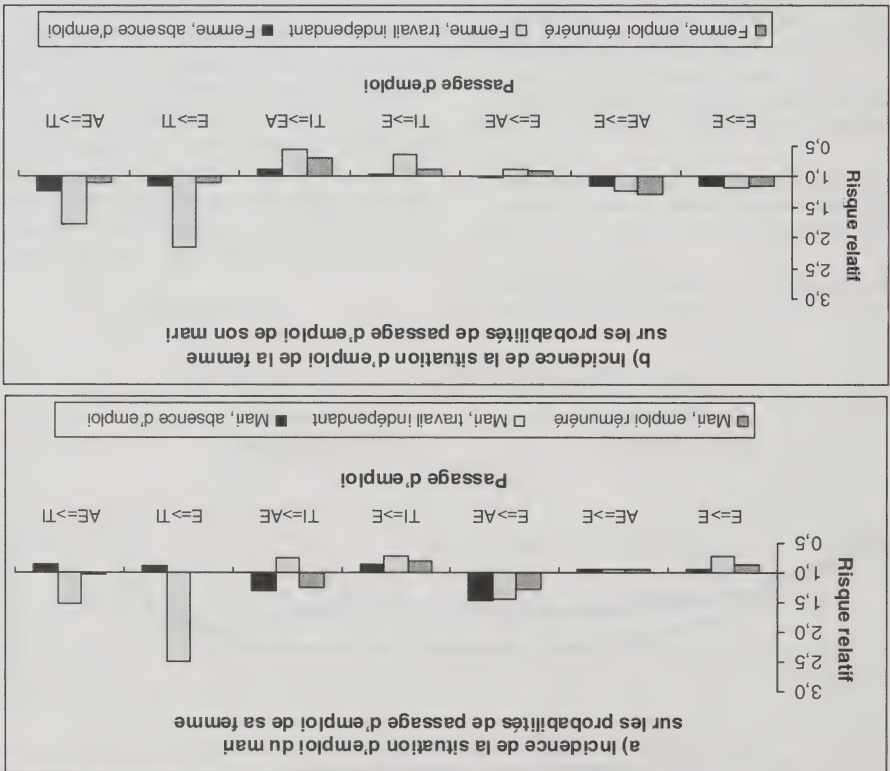
- ANDERSEN, P., et BORGAN, Ø. (1985). Counting process models for life history data: A review. *Scandinavian Journal of Statistics*, 12, 97-158.
- BORGAN, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics*, 11, 1-16.
- CORAK, M., et HEISZ, A. (1995). Guide de l'utilisateur la durée du chômage. Direction des études analytiques documents de recherche, No. 84, Statistique Canada.
- HECKMAN, J.J., et HONORE, B.O.E. (1989). The identifiability of the competing risks model. *Biometrika*, 76(2), 325-330.
- HOLFORD, T.R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics*, 36, 299-305.
- KINACK, M. (1991). Measuring data quality with longitudinal data. *1991 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 514-519.
- LAIRD, N., et OLIVIER, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76, 231-240.
- LAWLESS, J.F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82, 808-815.
- LEMAITRE, G. (1988). The measurement and analysis of gross flows. Document de travail, Labour and Household Surveys Analysis Division, Statistique Canada.
- LEMAITRE, G., PICOT, G., et MURRAY, S. (1992). Le roulement de la main-d'œuvre : une vue d'ensemble. *L'emploi et le revenu en perspective*, 4(2), Statistique Canada.
- LINDSEY, J.K. (1995). Fitting parametric counting processes by using log-linear models. *Applied Statistics*, 44, 201-212.
- PRENTICE, R.L., KALBFLEISCH, J.D., PETERSON, A.V., FLOURNOY, JR. N., FAREWELL, V.T. et BRESLOW, N.E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34, 541-554.
- ANDERSEN, P. (1985). Statistical models for longitudinal labor market data based on counting processes. Dans *Longitudinal Analysis of Labor Market Data*. (Eds. James J. Heckman et Burton Singer). Cambridge University Press, Cambridge.
- ALLOUM, A., et COMMENGES, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, 52, 512-524.

4. DES PARAMÈTRES ESTIMÉS AUX RÉSULTATS DE SIMULATION : UNE ILLUSTRATION

Notre exemple portant sur l'incidence de la situation d'emploi du conjoint fait voir la nécessité de tenir compte du contexte familial dans toute simulation de l'activité. Il est difficile dans le modèle LifePaths d'intégrer des relations de ce genre au cadre de simulation. Ainsi, dans une modélisation peu appropriée et peu précise de la scolarité sur les passages d'emploi, les conséquences iront des rapports directs scolarité-emploi à une chaîne d'effets indirects (relations entre scolarité et mariage, fécondité, migration interprovinciale, etc.). Ces effets seront repris pour le conjoint en simulation, comme nous l'avons indiqué. On peut facilement voir que, faute d'une spécification appropriée de ces relations et d'une précision acceptable des paramètres estimés, les biais se propageront à toutes sortes de résultats en simulation.

Nous en sommes venus à une validation générale des équations de probabilité d'emploi de LifePaths en comparant les rapports moyens annuels emploi/population en simulation aux estimations transversales directes de l'EPA. Nous avons tiré nos rapports emploi/population d'une population synthétique dont les membres ont suffisamment été exposés à un des sept types de probabilités de passage d'emploi dans chaque année de simulation. Nous les avons calculés à partir du nombre annuel d'années-personnes d'emploi de cette population. En d'autres termes, ils sont le résultat de la simulation des passages à l'emploi et hors de l'emploi. Nous avons nécessairement dû produire des distributions appropriées de covariables qui, à leur tour, ont déterminé les distributions de probabilité de transition d'emploi. Comme on peut le voir à la figure 8, LifePaths rend fidèlement compte de la structure par âge de l'emploi féminin tant en 1976 qu'en 2001, et donc de l'évolution considérable observée sur ce plan depuis 25 ans.

Figure 7. Incidence de la situation d'emploi du conjoint sur les probabilités de passage d'emploi



les gens qui étudient à plein temps seront plus actifs sur le marché du travail pendant la période estivale, alors que les congés de maternité que prennent les femmes qui ont un emploi pourraient largement être déterminés par le régime d'assurance-emploi. Voilà pourquoi LifePaths distingue les groupes suivants et modélise séparément leur activité :

- gens qui étudient à plein temps;
- gens qui viennent de faire des études complètes ou incomplètes et qui opèrent leur passage à l'emploi;
- femmes enceintes admissibles à un congé de maternité;
- gens dans la force de l'âge en matière d'emploi;
- travailleurs plus âgés en transition de retraite.

(14 équations au total) :

Il ne sera question ici que de l'estimation relative au quadrinme de ces groupes, c'est-à-dire aux gens qui, dans le modèle LifePaths, en sont à l'étape de « l'emploi de carrière » (la plus importante pour son incidence sur l'économie). Les caractéristiques des autres groupes sont décrites au site Web de Statistique Canada dont nous avons fait mention.

Dans son application dans LifePaths, notre modèle de probabilité emploie une forme logarithmique d'équation de régression pour les sept passages et les deux sexes

$$E(Y_{j,t+k}) \approx \hat{h}_{j,t+k} = \exp \left(\hat{g}_{m_{j,t+k}} + X_{j,t+k} \beta \right), \quad (4)$$

où $E()$ est l'opérateur de l'espérance, $g(m)$ l'opérateur spline logarithmique de la durée, X un vecteur de covariables à variation temporelle et β un vecteur de coefficients de régression. Le terme $g(m)$ correspond à une probabilité de référence de Weibull par morceaux, ce qui, dans notre spécification, distingue les risques de passage d'emploi entre les durées de moins et de plus d'un an. Les covariables X sont celles de l'âge et de la scolarité des gens, de la province de résidence, de la présence d'enfants par tranche d'âge, de la situation d'emploi du conjoint et de l'année et du mois civils avec les interactions de certaines de ces valeurs. Les estimations finales de β et $g(m)$ minimisent la somme des carrés d'écarts (3).

Dans la seule illustration des résultats détaillés que nous présenterons, nous tenons compte de l'incidence des situations d'emploi respectives du mari et de la femme sur les probabilités de passage chez l'un et l'autre. À la figure 7, nous comparons les estimations des coefficients des sept équations correspondant aux sept passages spécifiés. Les deux parties de cette figure présentent les équations respectives des hommes et des femmes. La catégorie « absence de conjoint » est notre catégorie de référence et la situation d'emploi du conjoint est « emploi rémunéré », « travail indépendant » ou « absence d'emploi ». Nous indiquons les coefficients estimés comme risques rapportés à ceux du groupe de référence. Après prise en compte des autres covariables, la probabilité de travail indépendant chez les femmes ayant un emploi et dont le mari est travailleur indépendant est d'environ 2,5 fois supérieure à la probabilité calculée pour leurs homologues sans conjoint (voir la barre la plus haute dans la partie supérieure de la figure).

La figure 7 indique que la présence même d'un conjoint peut être d'une incidence contraire chez les hommes et les femmes. Pour les deux sexes, les passages les plus fréquents sont $E => E$, $AE => AE$. Chez les femmes, les deux premières de ces passages sont moins probables si elles sont mariées que si elles sont célibataires; le passage à « absence d'emploi » est plus probable. (La présence d'enfants n'en est pas la raison, car elle est prise en compte dans d'autres termes de l'équation.) Chez les hommes, les tendances sont contraires. Ainsi, ces résultats semblent le reflet des rôles habituels des sexes, mais si nous tenons compte de l'ordre de grandeur de ces risques relatifs, nous n'avons pas l'impression que le partage des tâches entre hommes et femmes influe outre mesure sur les tendances une fois que les autres variables sont prises en compte.

Une autre tendance se remarque d'emblée à la figure 7. D'abord, les risques relatifs de passage au travail indépendant chez les gens dont le conjoint est travailleur indépendant sont les plus élevés pour tous les passages considérés. En second lieu, ces gens ont les risques relatifs les plus bas de passage hors de la situation de travailleur indépendant. Ainsi, cette situation semble être en renforcement entre mari-femme dans les familles. Ces observations vont dans le sens du constat du travail indépendant commun dans une entreprise familiale (dépanneurs, par exemple) ou de l'endogamie professionnelle (avocats qui marient quelque un de leur profession, par exemple).

probabilité. L'hypothèse selon laquelle les fonctions de probabilité sous-jacentes sont approximativement constantes par morceaux même directement à la somme des cartes de moyennes des fonctions de probabilité d'emploi (comme celles de la figure 3) en raccorderait implicitement toutes les indications disponibles sur les membres d'une cohorte définie dans les échantillons longitudinaux de l'EPA, c'est-à-dire en maximisant la vraisemblance (1), mais sans prendre de covariables en considération. Un tel traitement de la censure et de la troncature offre un exemple relativement simple des problèmes qui se posent si on le compare aux plans d'observation plus complexes envisagés par Alioum et Commenges (1996). Ce raccourciement implicite des indications se remarque dans la somme des cartes d'écarts (3) où il y a deux composantes, l'une qui est non nulle seulement aux passages observés et l'autre qui traduit les différences pondérées entre les événements et les probabilités respectivement cumulés (en cumulation sur l'ensemble des durées avant les événements ou les points de censure). Dans la mesure où les coupes transversales de l'EPA sont des échantillons représentatifs des semaines de référence, elles livrent collectivement une juste estimation du nombre de changements qui se produisent au cours de la « vie » d'une cohorte d'emploi. De même, dans les échantillons tirés de cohortes d'emploi, on peut s'attendre à trouver des périodes de répondants tronquées à gauche et à droite par lesquelles on pourrait imputer les périodes manquantes de troncature à gauche qui se terminent par un passage. C'est ainsi que la première composante de la somme des cartes d'écarts sera le reflet fidèle de la tendance des estimations de probabilité à devenir importantes dans des périodes où les changements observés sont fréquents. Quant à la seconde composante sommée sur tous les mois des enquêtes, sa valeur pourrait correspondre à celle que nous aurions pu obtenir s'il n'y avait eu aucune troncature à gauche. Pour des données aussi vastes que celles-là, la vraisemblance conditionnelle équivaut presque à une vraisemblance inconditionnelle.

3.2 Estimation d'équations de probabilité de passage d'emploi

Les tendances de passage d'emploi varient d'une manière significative selon les groupes démographiques. Ainsi,

devient une variable aléatoire de Poisson dont l'espérance est égale à la probabilité $h_{j,t+k}^{j,t+k}$, qu'on suppose constante par morceaux. Dans ce modèle, la contribution de i à la logvraisemblance sur n périodes (nous utilisons $h_{j,t+k}^{j,t+k} = f_j(m_{t+k}^{j,t+k}) / (1 - F_j(m_{t+k}^{j,t+k}))$) avec $f_j(m_{t+k}^{j,t+k})$ est approximativement :

$$\ln(L_{ij}) \approx \sum_{k=1}^K \left[Y_{j,t+k} \ln(h_{j,t+k}^{j,t+k}) - h_{j,t+k}^{j,t+k} \right]. \quad (2)$$

Il est de règle de prendre en compte un plan d'échantillonnage complexe par une « pseudovraisemblance » à laquelle est intégrée la pondération d'enquête. Maximiser la « pseudovraisemblance », c'est minimiser la somme pondérée des cartes d'écarts (termes représentant la différence entre les contributions estimées à la vraisemblance et leurs valeurs maximales possibles). Ainsi, la logvraisemblance conditionnelle d'échantillon entier pour le passage j peut se transformer en une somme pondérée de cartes d'écarts D_j (à noter que W vient de la pondération d'enquête et que, comme les passages sont normalement reconnus par comparaison des situations d'emploi entre les interviews, nous prenons les moyennes des valeurs consécutives de pondération d'enquête transversale pour obtenir W) :

$$D_j \approx -2 \left(\sum_{i=1}^n \sum_{k=1}^K W_{t+k}^{j,t+k} Y_{j,t+k} \ln(h_{j,t+k}^{j,t+k}) \right) + \left(\sum_{i=1}^n \sum_{k=1}^K W_{t+k}^{j,t+k} \left[Y_{j,t+k} - h_{j,t+k}^{j,t+k} \right] \right)^2. \quad (3)$$

Dans l'analyse de chaque passage du type j , nous considérons les autres événements (événements autres que le changement j qui se produisent dans la même population « à risque ») comme ceux que la somme des cartes d'écarts d'un ensemble de tels événements sera la somme de cartes d'écarts des composantes (si la probabilité générale est la somme des probabilités rivales, les risques en réalité peuvent être tenus pour indépendants (Prenhice et coll. 1978)).

Dans une justification plus directe de cette somme de cartes d'écarts, nous prenons un traitement de Poisson comme point de départ (Borgan 1984; Andersen 1985; Andersen et Borgan 1985; Lawless 1987) plutôt que de postuler des densités de durées latentes selon les événements ($f_j(m_{t+k}^{j,t+k})$). Dans ce cas, nous pouvons modéliser des dénombrements sur échantillon à variables multiples qui représentent le nombre de passages particuliers dans un intervalle temporel $[t_0, t]$. Les dénombrements sur échantillon représentés par les fonctions échelons de la figure 6 sont le pendant observable des fonctions cumulées de

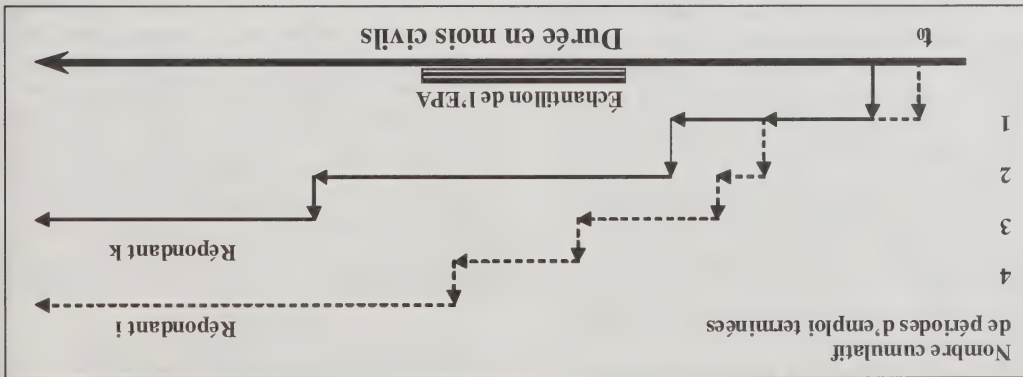


Figure 6. Changements récurrents et durées de période d'emploi observables par la fenêtre d'échantillonnage de l'EPA

($f_j(m_{t_i+k}|m_{t_i})$) ou d'une probabilité conditionnelle de survie – sans que s'opère précisément le passage f – par rapport à la durée observée ($(1 - F_j(m_{t_i+k}|m_{t_i}))$) selon qu'il y a eu ou non censure.

$$L_{t_i+t+k} = f_j(m_{t_i+k}|m_{t_i}) \left(1 - F_j(m_{t_i+k}|m_{t_i})\right)^{1-C_{f,t_i+k}} \\ = \frac{\left(1 - F_j(m_{t_i})\right)^{1-L_{t_i}}}{\left(1 - F_j(m_{t_i+k})\right)^{1-C_{f,t_i+k}}} \quad (1)$$

Dans ce calcul de vraisemblance, nous exploitons toutes les indications dont nous disposons sur le risque particulier

de passage f et pouvons tenir compte de l'effet de risques à droite s'ajoutant à la sortie de l'échantillon par renouvellement. On exprime fréquemment les problèmes de risques à différents niveaux en durées latentes, notamment en épidémiologie et en biostatistique, mais aussi en économique (Heckman et Honoré 1989, par exemple). Si on y a vu un traitement mathématiquement commode de la vraisemblance, on a

toutefois critiqué cette méthode en invoquant les hypothèses gratuites, l'absence d'interprétation objective et les problèmes d'identifiabilité (Prenette, Kalbfleisch, Peterson, Floumoy, Farewell et Breslow 1978).

On peut approcher les probabilités conditionnelles (1) par une équation de vraisemblance de Poisson (Holford 1980; Laird et Olivier 1981), reconnaissant aussi de ce fait le

caractère discontinu des données (les passages s'observent généralement dans le mois qui sépare deux interviews consécutives). On peut réexprimer l'équation (1) sous forme de variable binaire (X_{j,t_i+k}) qui représente la présence ou l'absence d'un passage dans un certain laps de temps (à noter que $X_{j,t_i+k} = 1 - C_{f,t_i+k}$). Dans ce traitement, X_{j,t_i+k}

La réunion des indications complètes et partielles livrées

par les données tronquées à gauche et à droite peut se représenter sous forme de probabilités conditionnelles (Wang 1991). Dans un cadre descriptif de risques niveaux, la

probabilité d'un passage d'emploi du type f du répondant i peut s'exprimer par la durée constatée k mois après que i a

changé d'emploi. Soit t_i désignant l'année et le mois de l'interview de l'EPA où la situation d'emploi en cours de t_i

a d'abord été observée (c'est souvent la première interview). D'après les renseignements réunis aux diverses interviews,

il est possible d'établir la durée de la période d'emploi ou d'absence d'emploi en cours (m_{t_i}). $m_{t_i+k} = m_{t_i} + k$ de-

signe alors la durée écoulée de la situation évaluée k mois après la première observation – nous supposons qu'aucun

changement ne s'est produit dans l'intervalle – et la probabilité d'un passage du type f (L_{f,t_i+k}) peut s'exprimer

par m_{t_i+k} . Voici les termes de la fonction de vraisemblance : densité de probabilité des durées jusqu'au pas-

sage du type f ($f_j(m_{t_i+k})$), probabilité cumulée correspon-

dante ($F_j(m_{t_i+k})$), variable binaire indiquant s'il y a eu censure ou non (C_{f,t_i+k}), variable binaire indiquant s'il y a eu troncature à gauche (LT_{t_i}). Il convient de noter que, dans le cadre descriptif de risques niveaux, la densité $f_j(m_{t_i+k})$ vise une variable latente, celle de la durée jusqu'au passage f précisément, et qu'on doit supposer qu'une telle densité existe pour chaque changement rival. La durée terminée (observée lorsqu'un passage s'opère) correspondra principalement au minimum des durées latentes en rivalité.

Pour tenir compte de la troncature à gauche, nous prenons la probabilité conditionnelle compte tenu de la durée d'abord observée (m_{t_i}). Il s'agit de la probabilité conditionnelle évaluée au moment où on observe un passage

3. UTILISATION DE MICRODONNÉES LONGITUDINALES DE L'EPA POUR LA MODÉLISATION DE L'ACTIVITÉ DANS LIFEPAATHS

Dans cette section, nous verrons comment les données de l'EPA peuvent être mises au service de la simulation de l'activité dans LifePaths. À l'heure actuelle, ce modèle range les situations d'emploi dans trois catégories, celles de l'emploi (E), du travail indépendant (TI) et de l'absence d'emploi (AE). Nous n'avons pas analysé les passages de chômage. (Le chômage est une situation complexe appelant un surcroît de questions de vérification et, comme nous l'avons signalé, la description des passages de chômage risque particulièrement d'être entachée d'une erreur de réponse).

Il y a six changements possibles de situation d'emploi (voir la figure 5). LifePaths les modélise tous, tout comme les passages qui ne semblent pas donner lieu à une interruption d'emploi (ce que nous désignerons ici par $E \Rightarrow E$). Par les microdonnées de l'EPA, nous avons estimé des équations de probabilité pour chacun de ces sept passages. Les coefficients estimés de ces équations deviennent les paramètres du module « Career Work » de LifePaths. Nous traiterons de certaines questions techniques tenant aux limites des données de l'EPA, puis donnerons des exemples de nos résultats d'estimation d'abord et de nos résultats de simulation ensuite.

Le caractère fragmentaire de ces données est un défi pour l'analyste. Il se pose l'importante question de savoir si des biais découlent inévitablement de ce caractère des données. Nous répondons en général qu'il est possible de tenir compte des limites en question et de prévenir les causes de biais par une analyse soignée.

3.1 Censure et/ou troncation de relevés d'activité

Pour l'analyste de ces données, un sujet d'inquiétude est l'absence de données rétrospectives en dehors des données

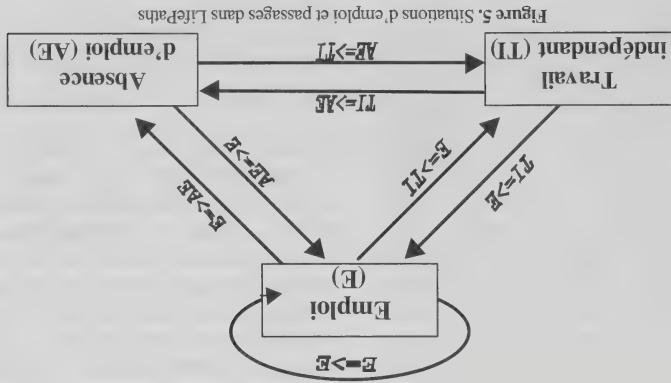
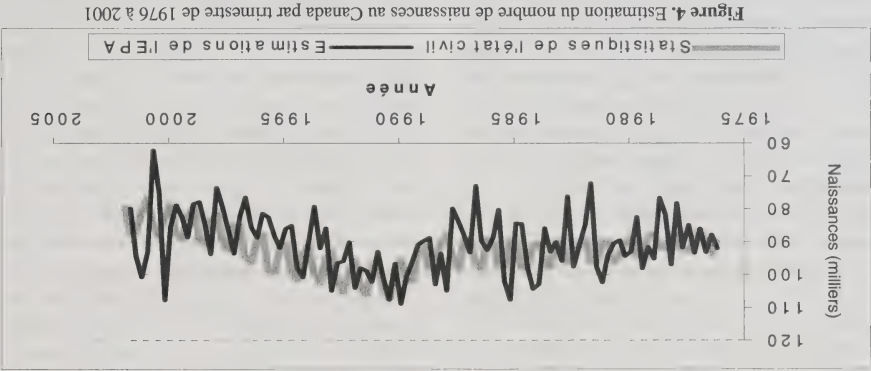
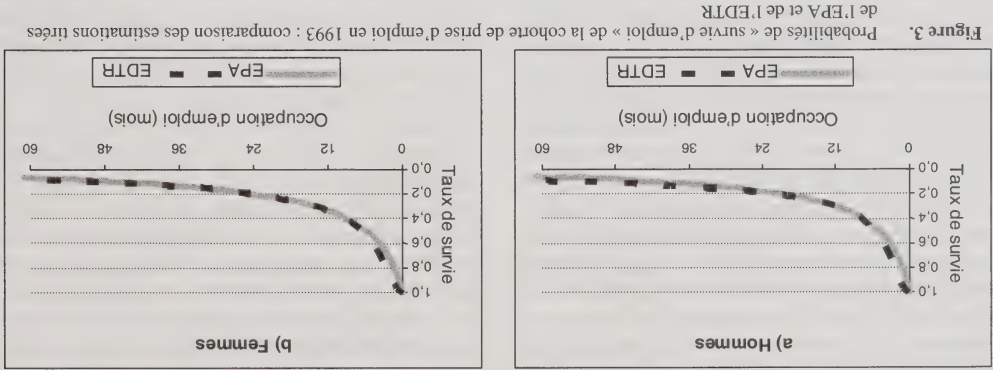


Figure 5. Situations d'emploi et passages dans LifePaths

de durée de la période d'emploi *en cours*. On pourrait considérer que les relevés individuels d'activité consistent en une succession (largement inobservée) de situations contingentes d'emploi (voir la figure 6) où les passages traduisent le cheminement de carrière. Ainsi, si on n'a que la description des passages observables dans l'EPA, les taux de passage que nous pourrions estimer comporteront nécessairement une mise en commun des données de répondants ayant eu des carrières très différentes. En revanche, des enquêtes à échantillon permanent comme l'EDTR recueillent des données rétrospectives à la première interview. Bien que limitée, cette information permet au moins d'évaluer l'expérience des intéressés, c'est-à-dire les longues interruptions d'emploi ou les périodes de travail à temps partiel dans le passé.

Un autre sujet d'inquiétude qu'illustre la figure 6 est que les périodes d'emploi selon l'EPA peuvent être tronquées à gauche et/ou à droite. S'il y a censure à droite, c'est qu'une période d'emploi cesse d'être observée ou que le répondant cesse d'être « à risque » sans que s'opère un passage de la nature visée. Cela peut se produire (1) parce que le ménage répondant est sorti par renouvellement de l'échantillon de l'EPA avant qu'un passage ne s'opère ou (2) qu'il s'est opéré un passage qui n'est pas de la nature visée. De même, les données seront fréquemment tronquées à gauche. Dans ce cas, le début d'une période reste inobservé parce qu'un passage s'est opéré avant que le ménage du répondant n'entre par renouvellement dans l'échantillon de l'EPA. (Ces troncations à gauche diffèrent des troncations à droite, car les répondants fournissent les indications nécessaires à la détermination de la durée écoulée de la période en cours au moment de la première interview.) Comme la censure et la troncation sont généralement indépendantes des changements de situation d'emploi, elles ne devraient engendrer aucun biais dans l'estimation des probabilités de passage si on tient bien compte du phénomène dans la fonction de vraisemblance.



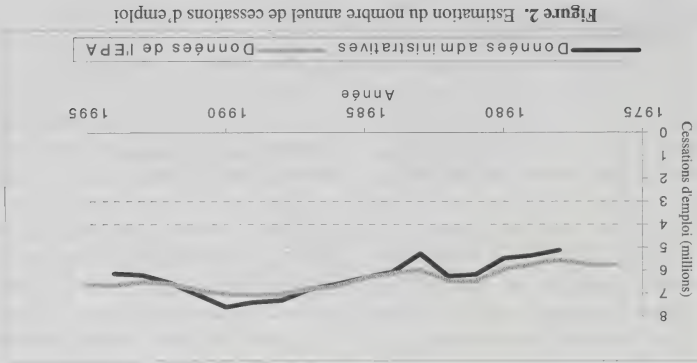
Nous avons estimé les probabilités de « survie d'emploi » à l'aide des données de l'EPA par le produit chaîne des taux moyens de maintien tirés des taux mensuels de cessation d'emploi principal de la période 1993-1998. Nous avons estimé de même ces probabilités par les données de l'EDTR en nous reportant aux données déclarées d'occupation et aux dates de fin d'emploi. Les deux courbes de survie ont la même forme caractéristique, indiquant une déperdition relativement importante des emplois de moins d'un an et une déperdition bien moindre des emplois de un à cinq ans. Les estimations des emplois d'environ six mois ou moins sont en divergence, ce qui peut s'expliquer par la période de rappel d'un an dans les interviews de l'EDTR et par la restriction de cessation d'emploi principal dans celles de l'EPA. Pour les emplois de jusqu'à cinq ans, les estimations des deux enquêtes n'en sont pas moins des plus convergentes. Avec les données disponibles de l'EPA, il est possible de suivre certaines cohortes d'emploi jusqu'à 25 ans après la prise d'emploi.

Dans un dernier exemple d'exploitation longitudinale efficace des données de l'EPA, nous mettons en évidence des microdonnées longitudinales utiles. Nous avons comparé le nombre d'enfants âgés de moins d'un an selon les déclarations des chefs féminins de famille économique ou des conjoints des chefs masculins. Un bébé nouvellement déclaré par une femme de 15 à 50 ans correspond probablement à une naissance. Pour une comparaison mensuelle le nombre d'enfants âgés de moins d'un an selon les déclarations des chefs féminins de famille économique ou des conjoints des chefs masculins, nous tenons compte, par des corrections simples, de la proportion de naissances chez les autres femmes des familles économiques (mères adolescentes demeurant avec leurs parents, par exemple), ainsi que des naissances au Yukon, dans les Territoires du Nord-Ouest et au Nunavut. Le rapprochement des estimations mensuelles des naissances de l'EPA et des chiffres correspondants de dénombrement des naissances dans les statistiques de l'état civil (figure 4) démontre que les premières dégageant les tendances à long terme de la fécondité, tout en saisissant une partie des variations mensuelles des naissances. Ensemble, ces trois exemples indiquent que, à condition de bien prendre en compte les aspects de l'observation et des concepts d'enquête et de l'erreur possible de réponse, l'EPA peut nous livrer des microdonnées longitudinales utiles.

Par le passé, on s'est reporté aux données transversales de l'EPA pour estimer les fréquences d'embauchages et de cessation d'emploi à intervalles mensuels (Lemaitre, Picot et Murray 1992). Dans ce dernier cas, on a directement observé les embauchages par la fréquence des occupations d'emploi déclarées d'au plus un mois et déterminé par voie résiduelle les cessations à l'aide d'estimations globales des changements d'emploi et d'estimations des embauchages. Ces données ont aussi servi à calculer et à comparer les statistiques de durée d'emploi de cohortes synthétiques. Ainsi, Corak et Heisz (1995) représentent l'expérience d'une cohorte hypothétique par les taux de maintien de l'emploi d'un seul intervalle temporel. Ils ont tiré ces taux du dénombrement des répondants de l'EPA ayant un emploi qui ont indiqué une occupation « *t* » pour le mois « *m* » et de ceux qui ont indiqué une occupation « *t + 1* » le mois suivant. De tels usages de données transversales accusent certaines limites. Mentionnons en particulier que, comme les mouvements des gens ne sont pas directement observés, les états de destination sont inconnus. (S'il est possible d'estimer une proportion en cessation d'emploi, il est impossible d'estimer la proportion de ceux qui sont devenus chômeurs au lieu de passer à l'inactivité ou de prendre sur les proportions d'emplois d'une certaine durée – peut servir d'indice sensible aux variations des conditions du marché du travail.)

2.1 Validation de principe : exemples choisis de validation de données longitudinales

Les données de l'EPA n'étaient pas destinées à une exploitation longitudinale et celle-ci peut poser des problèmes (Stansy 1986; Lemaître 1988; Kinack 1991). Il importe donc de vérifier pour chaque analyse s'il est possible de tirer des estimations valides de la comparaisons de réponses longitudinales de mois en mois. Nous citerons



La figure 3 valide à son tour la dynamique de l'emploi par une comparaison de « survie d'emploi » chez les hommes et les femmes ayant pris un emploi en 1993 selon les estimations de l'EPA et de l'EDTR. (À noter que 1993 est la première année de collecte des données de l'EDTR).

(Statistique Canada 1998).

Comme on peut le voir, le nombre de passages est déterminé par la comparaison mensuelle des données de l'EPA et correspond de près au nombre établi par les données RE. Il subsiste des différences entre les deux séries. Certaines pourraient tenir à des différences d'observation entre les données de l'EPA et les données administratives, ainsi qu'aux changements périodiquement apportés au plan de sondage ou au questionnaire de cette enquête. Il peut aussi se présenter des différences parce que, dans notre dénombrement par les données de l'EPA, on ne tient pas compte des cessations en cumulé d'emplois où le travailleur a conservé au moins un de ses emplois (les dénombrements visent seulement les changements d'emploi principal). Nous n'en considérons pas moins que la concordance des deux séries est suffisante pour justifier que l'on puisse l'analyser des microdonnées de l'EPA. Pour les deux sources de données, le taux annuel de cessation d'emploi est élevé. Pour les données RE de la période 1978-1995, le taux annuel moyen chez les hommes est de plus de 38 % du nombre annuel d'emplois-personnes. Une plus ample analyse des microdonnées de l'EPA peut nous éclairer sur

Ce n'est pas la première fois qu'on tente d'exploiter longitudinalement les données de l'EPA. Slansky (1986) et Lemaitre (1988) ont étudié les erreurs d'estimation de « flux bruts » de situation d'activité (*emploi*, *chômage* et *inactivité*) à intervalles mensuels. Lemaitre a constaté que des problèmes se posaient à la fois par erreur de réponse et parce que les concepts de l'Enquête sur la population active (plan de sondage transversal tendent à créer des « flux » lorsqu'on relie les réponses de mois consécutifs. (Comme exemple, on peut citer le traitement des *travailleurs en disponibilité* et des *travailleurs autonomes hors entreprise*. Il n'en conclut pas moins que les données administratives montrent que ce ne sont pas tous les sous-groupes de changement de situation d'emploi qui sont largement surestimés. Pour sa part, Kinack (1991) a examiné la cohérence longitudinale des réponses aux questions sur la recherche d'emploi par lesquelles on distingue les catégories « *chômage* » et « *inactivité* ». Il a conclu à une nette incohérence, surtout dans le cas des réponses par procuration provenant de différents répondants. Ces études nous indiquent que, en s'attachant aux passages entre les catégories « *emploi* » et « *absence d'emploi* » (c'est-à-dire sans distinguer les *chômeurs des inactifs*), on peut plus facilement atténuer l'incidence de l'erreur de réponse.

À la figure 1, on illustre certaines caractéristiques des données de l'EPA reformées en fragments longitudinaux, nous examinons les changements de situation d'emploi des membres de la cohorte qui ont pris un emploi en janvier 1976. Les répondants de cette cohorte qui sont entrés dans l'échantillon par le groupe de renouvellement I apportent des données les six premiers mois, soit de janvier (début de l'emploi) à juin (sortie de l'échantillon) 1976. Pour les membres du groupe de renouvellement 2, les données longitudinales sur six mois sont décalées à droite d'un mois (entrée et sortie décalées d'un mois par rapport à celles du groupe de renouvellement I). Il en va de même des groupes de renouvellement ultérieurs. Ainsi, on peut constater les données longitudinales de l'EPA comme une combinaison d'ensembles de données longitudinales en chevauchement où les membres d'un même groupe de renouvellement constituent un échantillon permanent (panel) au sens habituel du terme.

Il est possible de réunir les fragments successifs sur six mois de données longitudinales de l'EPA pour livrer des estimations cumulatives successives de la déperdition d'une cohorte initiale d'emplois, ainsi que pour reconnaître de

Groupe	Bar 1 (0-2)	Bar 2 (1-2)	Bar 3 (2-2)
Répondant du groupe de renouvellement 1	0	1	2
Répondant du groupe de renouvellement 2	1	2	2
Répondant du groupe de renouvellement 3	2	2	2
Répondant du groupe de renouvellement 4	2	2	2
Répondant du groupe de renouvellement 5	2	2	2
Répondant du groupe de renouvellement 6	2	2	2

Janv. 1976 Févr. 1976 Mars

Légende

=====	Emplois entrants en janvier 1976
.....	Absence d'emploi
=====	Emplois ultérieurs

Les chiffres sont ceux de l'occupation des emplois (en mois). Les flèches indiquent le maintien de la situation au moment de la sortie de l'échantillon de l'EPA.

Analyse longitudinale des données de l'Enquête sur la population active

GEOFF ROWE et HUAN NGUYEN¹

RÉSUMÉ

Au Canada, l'Enquête sur la population active (EPA) n'a pas au départ de caractère longitudinal, mais comme les ménages répondants demeurent normalement dans l'échantillon six mois de suite, il est possible de reconstituer des fragments longitudinaux sur six mois à partir des enregistrements mensuels des membres des ménages. De telles microdonnées à analyser par mois la dynamique du marché du travail, et ce, sur des périodes relativement longues de 25 ans et plus.

Nous employons ces données pour estimer des fonctions de probabilité décrivant les passages entre les situations d'emploi, à savoir le *travail indépendant*, le *travail rémunéré* et l'*absence d'emploi*. Avec les données sur l'occupation des emplois et le dernier jour travaillé des gens qui n'ont pas d'emploi, jointes aux données sur la date de réponse à l'enquête, on peut élaborer des modèles comportant des termes de saisonnalité et de cycle macroéconomique, ainsi que de durée de dépendance pour chaque type de passage. Ajoutons que les données de l'EPA permettent d'inclure des variables de l'activité du conjoint et de la composition de la famille dans les modèles de probabilité comme covariables à variation temporelle. Les équations estimées de probabilité ont été intégrées au modèle de microsimulation LifePaths. Dans ce cadre, nous avons pu par ces équations, simuler l'activité à vie de cohortes de naissances passées, présentes et futures. Nous avons validé les résultats de cette simulation par rapprochement avec les profils d'âge de la période 1976-2001 pour les rapports emploi/population de l'EPA.

MOTS CLÉS : Microsimulation; censure; troncation; dynamique de l'emploi.

1. INTRODUCTION

Ces dernières années, on convient de plus en plus de l'importance d'une étude de la dynamique du marché du travail à l'aide de (micro)données individuelles. C'est dans ce but qu'on a mis au point de nouvelles enquêtes à échantillon permanent comme l'Enquête sur la dynamique du travail et du revenu ou EDTR (Statistique Canada 1998). Il reste que les données de l'EPA dont nous disposons déjà (Statistique Canada 2002) représentent une source de données chronologiques pour ainsi dire inexploitées sous la forme de nombreux relevés chronologiques fragmentaires. Vues sous l'angle habituel, les données forment actuellement une série chronologique de plus de 300 enquêtes transversales réalisées à intervalles mensuels pendant plus de 25 ans. D'un point de vue longitudinal cependant, il s'agit d'environ 6,5 millions de relevés chronologiques fragmentaires par intervalles temporels en chevauchement depuis un quart de siècle et d'un total de plus de 34 millions de mois-personnes d'observation.

L'analyse que nous évoquerons vise précisément à l'élaboration de modèles de probabilité à intégrer à LifePaths (Statistique Canada 2001), modèle de microsimulation de la population canadienne. Le lecteur peut se renseigner plus en détail sur le modèle LifePaths au site Web de Statistique Canada www.statcan.ca/francais/spssd/index.htm.

Notre exposé est ainsi structuré : à la section 2, nous réorganisons certaines caractéristiques des données de l'EPA où sont comparées les estimations du fichier longitudinal obtenu et les estimations correspondantes puisées à l'activité dans le modèle LifePaths; il est ensuite question du recours aux microdonnées de l'EPA pour l'estimation de l'emploi; enfin, nous illustrons par quelques exemples les résultats d'estimation et une validation des simulations LifePaths faisant appel aux équations de probabilité.

2. DONNÉES LONGITUDINALES DE L'EPA : TRAITS DISTINCTIFS ET VALIDATION DE PRINCIPE

Nous avons élaboré une version longitudinale des données de l'EPA par chaînage des enregistrements mensuels des divers répondants dans un fichier contenant un enregistrement normalement dans l'échantillon six mois de suite, il est possible d'obtenir des relevés sur six mois pour la plupart des répondants. Ces relevés ne sont pas en soi assez longs pour la plupart des analyses longitudinales, mais

¹ Geoff Rowe et Huan Nguyen, Division de l'analyse socioéconomique et de la modélisation, Secleur de l'analyse et du développement, Statistique Canada, Ottawa, (Ontario), Canada, K1A 0T6.

REMERCIEMENTS

Nous tenons à remercier deux examinateurs anonymes de leurs suggestions, qui nous ont permis d'améliorer sensiblement notre étude.

BIBLIOGRAPHIE

- GIBBONS, R.D., et HEBBEKER, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, 53, 1527-1537.
- GOLDFELD, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 73, 43-56.
- GRABARD, B.I., et KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5, 263-281.
- HEBEKER, D., et GIBBONS, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50, 933-944.
- KORN, E.L., et GRABARD, B.I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society B*, 65(1), 175-190.
- PFEFFERMANN, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, 61(2), 317-337
- PFEFFERMANN, D., SKINNER, C.J., HOLMES, D.J., GOLDSCHMIDT, H., et RASBACH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*, 60(1), 23-40.
- RASBACH, J., BROWN, W., GOLDSCHMIDT, H., YANG, M., PLEWIS, I., HEALY, M., WOODHOUSE, G., et DRAPER, D. (1999). A users guide to MLWIN. London: Multilevel models project, Institute of Education, University of London.
- RENNARD, D., et MOLENBERGHS, G. (2002). Multilevel modelling of complex survey data. Dans *Topics in modelling of clustered data* (Ed. M. Aerts M). London: Chapman and Hall, 263-272.
- SÄRNDAHL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SAS INSTITUTE (1999). SAS/STAT User's Guide Version 8. Cary: SAS Institute Inc.
- SHAO, J., et TU, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer-Verlag.
- SINGH, A.C., FOLSOM, R.E. et VAISH, A.K. (2002). A hierarchical Bayes generalization of the Fay-Herriot method to unit level nonlinear mixed models for small area estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, New York, le 10-13 Août, 3258-3263.
- SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt et T.M.F. Smith), Chichester: Wiley, 59-87.

servant à ajuster les modèles non linéaires à effets aléatoires au moyen d'une version adaptée de la quadrature gaussienne. La procédure NLMIXED ne comporte pas d'options permettant une estimation du PMV, mais il est possible d'incorporer les coefficients de pondération à la vraisemblance en utilisant différentes astuces en ce qui touche les coefficients de pondération des niveaux 1 et 2. Pour incorporer des coefficients de pondération relatifs au niveau 1, il faut tirer parti de l'option permettant de formuler l'expression de la vraisemblance conditionnelle du modèle; il suffit ensuite de formuler l'expression $w_{ij} \log L_{ij}(\theta|u)$ (soit sous forme d'énoncés de programmation du SAS (se reporter à la section 3.1). Les coefficients de pondération relatifs au niveau 2 peuvent pour leur part être incorporés à la vraisemblance au moyen de l'énoncé `replicate`. Malheureusement, cet énoncé sert uniquement pour les coefficients de pondération qui sont des entiers, de façon à éviter les approximations trop imprécises. Il est donc conseillé de procéder de la manière suivante : a) multiplier tous les coefficients de pondération du niveau 2 par une constante arbitraire k (égale à 10 000 dans notre application); b) incorporer le nombre entier des coefficients de pondération majorés à la vraisemblance à l'aide de l'énoncé `replicate`; c) multiplier la matrice de covariance estimative par k au moyen de l'option `cfactor`. Cette façon de faire repose sur le fait que la multiplication des coefficients de pondération du niveau 2 par une constante a pour seul effet de majorer la matrice d'information en proportion de cette constante, les estimations demeurant ainsi inchangées. De toute manière, lorsque l'on utilise la méthode d'estimation pondérée pour effectuer des ajustements au titre d'un plan informatif, la matrice de covariance estimative des paramètres estimatifs n'est pas fiable.

Voici le code SAS, les symboles /* et */ délimitant les commentaires :

```
proc nlmixed data=dataname gpoints=10
/* cfactor=10,000;
matrice de covariance estimative des
paramètres estimatifs */
parms b0=0 sd=0.5; /* valeurs initiales */
bounds sd >= 0;
eta=b0+randef*sd;
if (yobs=1) then z=probnorm(eta);
else if (yobs=0) then z=-1-probnorm(eta);
if (z >= 1e-8) then ll=log(z); else ll=-1e10;
/* de façon à éviter les problèmes d'ordre
numérique si z devient trop petit */
ll=ll*w2; /* inclusion de coefficients de
pondération au niveau 1 */
model yobs=general(1,1);
random randef~normal(0,1 subject=;
/* j est l'identification de grappe */
w2; /* incorporation des
coefficients de pondération du niveau 2
(repliés seulement) */
ods output
ConvergenceStatus=cs;
run;
```


Tableau 5

Simulation – Ecart types des estimateurs ponctuels pondérés de plans informatifs aux deux niveaux (1 000 répétitions pour (à partir de 200 échantillons bootstrap dans chaque cas) dans le cas type de deuxième niveau, et estimations bootstrap correspondantes avec mise à l'échelle de l'ordonnée à l'origine ainsi que de l'écart

Plan d'échantillonnage informatif aux deux niveaux	β_0		ω_0	
	Simul.	Estim. relative	Simul.	Estim. relative
Taille fixe, $n_j = 38$	0,185	0,175	-5,4%	0,124
Taille prop., $n_j = 0,4N_j$	0,183	0,173	-5,5%	0,140
Taille fixe, $n_j = 9$	0,2	0,167	-16,5%	0,234
Taille prop., $n_j = 0,1N_j$	0,195	0,173	-11,3%	0,247
				117,8%

5. OBSERVATIONS FINALES

En raison du recours fréquent à des modèles binaires et ordinaux multivariés dans de nombreux domaines, notre étude a été consacrée aux effets des plans d'échantillonnage complexes sur l'ajustement de ce genre de modèle. Dans la présente étude, des simulations ont révélé qu'un plan d'échantillonnage complexe à deux degrés engendre un biais touchant l'ajustement d'un modèle binaire simple à l'ordonnée à l'origine aléatoire lorsque les probabilités de sélection des groupes ou des sujets dépendent des termes aléatoires du modèle. L'étude de simulation nous a également appris que, dans de telles situations, la procédure d'estimation à pondération probabiliste (PMV) décrite ici permet de réduire efficacement le biais. Cette procédure est facile à appliquer dans le SAS. Notamment, l'estimateur pondéré avec mise à l'échelle a produit un biais faible tant pour les paramètres fixes que pour les paramètres aléatoires, la variance d'échantillonnage n'augmentant que légèrement. Même lorsque la pondération s'avère inopérante, la perte d'efficacité attribuable à l'inclusion de coefficients de pondération avec mise à l'échelle apparaît très faible. L'application de la méthodologie proposée à des exemples réels exigera une stratégie opérationnelle fondée sur l'information disponible à propos du plan d'échantillonnage. On peut envisager deux cas limites : a) à chaque degré du plan d'échantillonnage, les probabilités d'inclusion et les ajustements au titre de la stratification et de la non-réponse sont connus avec exactitude; b) l'information disponible se limite aux coefficients de pondération finaux d'ensemble, y compris des ajustements au titre de la stratification a posteriori et de la non-réponse. Dans le cas a), les coefficients de pondération peuvent être calculés à chaque degré d'échantillonnage sous forme de valeurs inverses du produit des probabilités de sélection dans l'échantillon et des probabilités de réponse qui en

Nous présentons le code SAS utilisé pour appliquer les estimateurs à pondération probabiliste (PMV) décrits dans notre étude. L'élément essentiel du code est la procédure NL MIXED du SAS; il s'agit d'une procédure générale

ANNEXE A

Une autre question non résolue concerne le choix de la méthode de mise à l'échelle la plus efficace pour réduire le biais au niveau des estimateurs des composantes de la variance lorsque l'échantillon est de petite taille. La méthode du PMV décrite dans la présente étude est de nature tout à fait générale, et la technique d'estimation reposant sur la procédure NL MIXED du SAS est facile à généraliser en vue de l'appliquer à d'autres modèles non linéaires. Il serait donc intéressant d'évaluer les résultats de la méthode avec des modèles autres que les modèles binaires à l'ordonnée à l'origine aléatoire examinés ici.

Un autre question non résolue concerne le choix de la méthode de mise à l'échelle la plus efficace pour réduire le biais au niveau des estimateurs des composantes de la variance lorsque l'échantillon est de petite taille. La méthode du PMV décrite dans la présente étude est de nature tout à fait générale, et la technique d'estimation reposant sur la procédure NL MIXED du SAS est facile à généraliser en vue de l'appliquer à d'autres modèles non linéaires. Il serait donc intéressant d'évaluer les résultats de la méthode avec des modèles autres que les modèles binaires à l'ordonnée à l'origine aléatoire examinés ici.

Un des inconvénients de l'estimation à pondération probabiliste est la nécessité de prévoir des procédures spéciales pour estimer la variabilité des estimateurs. Dans l'application exposée ici, nous avons adopté une technique bootstrap simple sur le plan tant de la conceptualisation que de la programmation, mais qui exige une certaine somme de calculs. Si l'on se fie à notre étude de simulation limitée, les résultats obtenus sont bons uniquement dans le cas de grappes et d'échantillons de grande taille, toutefois, il faudrait procéder à davantage de simulations pour comprendre à fond le comportement de l'estimateur

Un des inconvénients de l'estimation à pondération probabiliste est la nécessité de prévoir des procédures spéciales pour estimer la variabilité des estimateurs. Dans l'application exposée ici, nous avons adopté une technique bootstrap simple sur le plan tant de la conceptualisation que de la programmation, mais qui exige une certaine somme de calculs. Si l'on se fie à notre étude de simulation limitée, les résultats obtenus sont bons uniquement dans le cas de grappes et d'échantillons de grande taille, toutefois, il faudrait procéder à davantage de simulations pour comprendre à fond le comportement de l'estimateur

Dans le cas b), le manque d'information est d'une importance critique car, même en l'absence de non-réponse et de stratification a posteriori, il n'est pas possible de dissocier les coefficients de pondération au niveau de la grappe et au niveau du sujet (conditionnels), du moins pas sans hypothèses solides. De ce fait, on ne peut effectuer une estimation pondérée.

Entre ces deux cas limites, il existe de nombreuses situations possibles qui exigent des solutions particulières. Par exemple, il arrivera souvent que le chercheur connaisse les probabilités d'inclusion au niveau de la grappe (π_j) et les coefficients de pondération finaux d'ensemble au niveau du sujet (w_j), qui comprennent également des ajustements au titre de la stratification a posteriori et de la non-réponse. Lorsque la stratification a posteriori et la non-réponse ont une incidence uniquement au niveau du sujet, les coefficients de pondération au niveau du sujet (conditionnels) peuvent être calculés ainsi : $w_{ij}^* = w_j \cdot \pi_j$. Kom et Graubard (2003) décrivent une situation plus complexe.

Entre ces deux cas limites, il existe de nombreuses situations possibles qui exigent des solutions particulières. Par exemple, il arrivera souvent que le chercheur connaisse les probabilités d'inclusion au niveau de la grappe (π_j) et les coefficients de pondération finaux d'ensemble au niveau du sujet (w_j), qui comprennent également des ajustements au titre de la stratification a posteriori et de la non-réponse. Lorsque la stratification a posteriori et la non-réponse ont une incidence uniquement au niveau du sujet, les coefficients de pondération au niveau du sujet (conditionnels) peuvent être calculés ainsi : $w_{ij}^* = w_j \cdot \pi_j$. Kom et Graubard (2003) décrivent une situation plus complexe.

estimateurs pondérés pour effectuer des ajustements et qu'un plan informatif est utilisé. Par exemple, l'erreur type d'estimateur entièrement pondéré avec ω_0 est égale à $n_j = 0,4N_j$ est égale à 0,109 pour β_0 et 0,089 pour ω_0 (comparativement à des valeurs de Monte Carlo de 0,183 et de 0,130, respectivement). On constate des biais par défaut de simulations avec les autres tailles d'échantillon, d'où la nécessité d'élaborer un autre estimateur de la variance.

La procédure bootstrap décrite à la section 3.4 a été utilisée pour estimer les écarts types d'échantillonnage des estimateurs ponctuels pondérés de β_0 et de ω_0 . Nous avons limité l'analyse à l'estimateur entièrement pondéré avec mise à l'échelle et aux plans informatifs aux deux niveaux. Pour réduire la somme de calculs requise, nous avons appliqué une procédure bootstrap qui ne tient pas compte des étapes relatives aux unités élémentaires – autrement dit, seules les grappes font l'objet d'un rééchantillonnage. Cette procédure devrait donner des résultats suffisamment exacts, étant donné la faible fraction d'échantillonnage (35/300) des grappes (se reporter à la section 3.4). Il y a eu au moyen du plan à deux degrés décrit à la section 4.1, et 200 échantillons bootstrap sont sélectionnés. Le tableau 5 expose, relativement à chaque paramètre, l'erreur type de Monte Carlo de la distribution d'échantillon de l'estimateur pondéré mise à l'échelle pour 1 000 répétitions du plan complexe (se reporter aux tableaux 1 et 2), ainsi que l'estimation bootstrap moyenne correspondante et le biais relatif.

Étant donné que les calculs demandent beaucoup de biais relatif.

temps, nous avons limité notre expérience à une procédure bootstrap partielle, reposant sur 200 échantillons boot-strap seulement. Des travaux supplémentaires sont requis pour calibrer le nombre d'échantillons bootstrap et pour étudier des variantes éventuelles de la méthode. Néanmoins, les chiffres présents au tableau 5 donnent certaines indications concernant le comportement des estimateurs bootstrap.

L'estimation de l'écart type d'échantillonnage donne de meilleurs résultats pour l'estimateur de β_0 que pour celui de particulier dans le cas de ω_0 : lorsque la grappe est de petite taille ($n_j = 9$ et $n_j = 0,1N_j$), l'estimation bootstrap n'est absolument pas fiable; à l'opposé, lorsque la grappe est de grande taille ($n_j = 38$ et $n_j = 0,4N_j$), les résultats sont passablement bons, étant donné que, tant pour β_0 que pour ω_0 , la méthode bootstrap donne une légère sous-estimation de la variance réelle. Il faut ajouter que les mauvais résultats produits par l'estimateur de la variance dans le cas de ω_0 n'ont pas une importance critique, étant donné que, de toute manière, il n'est généralement pas recommandé d'opter pour des tests de Wald à l'égard des paramètres de la variance en temps ordinaire.

4.2.6 Estimation de la variance bootstrap

ω_0 soulève les problèmes observés dans le modèle linéaire lors de l'estimation des deux composantes de la variance. Les simulations montrent que les résultats de l'estimateur pondéré avec mise à l'échelle de ω_0 ne sont pas satisfaisants. L'une des manières pouvant permettre d'obtenir de meilleurs résultats avec l'estimateur consiste à utiliser une autre méthode de mise à l'échelle. Korn et Graubard (2003) se sont penchés sur toute la question de la mise à l'échelle dans le contexte du modèle linéaire et ont formulé une mise en garde : selon eux, la méthode de mise à l'échelle utilisée ici (la deuxième méthode dont traitent Pfeffermann et coll., 1998) pourrait produire un biais important dans le cas de certains plans, et ce, même si les grappes et les échantillons à l'intérieur de ces grappes sont de grande taille. Pour nous faire une idée de l'importance de ce biais, nous avons mené une petite étude de simulation en reprenant le scénario défavorable énoncé par Korn et Graubard (2003), soit un échantillon aléatoire simple de grappes avec une population de taille égale, et, à l'intérieur de chaque grappe échantillonnée, un échantillon aléatoire simple de particuliers de taille $2m$ ou $m/2$ pour une valeur fixe de m , selon que la variabilité observée à l'égard des particuliers dans les grappes a tendance à être grande ou petite, respectivement. Les coefficients de pondération mis à l'échelle au niveau du sujet sont tous égaux à 1, ce qui rend le processus de pondération inefficace. En conséquence, dans le modèle linéaire des composantes de la variance, le biais de la variance interne sera élevé. Pour connaître la forme que prend ce comportement dans le modèle binaire de l'ordonnée à l'origine aléatoire, nous avons simulé 1 000 ensembles de données comportant 80 grappes dont la taille était de 36 ou de 9, selon que la variance binaire des réponses de la grappe est supérieure ou inférieure à la médiane, respectivement. Lorsque l'on utilise le même modèle de super-population que celui employé dans les simulations principales, les moyennes (et les écarts types) sont de -0,003 (0,098) pour β_0 et de 0,451 (0,144) pour ω_0 . La variance de nos simulations. Si l'estimation des composantes de la variance est de première importance, il faudra absolument améliorer la méthode employée, ce qui exigera des recherches plus approfondies.

La matrice de la covariance estimative applicable au paramètre estimé obtenu par l'inversion de la matrice d'information (produite par défaut à l'aide de la procédure NLMIXED) n'est pas fiable lorsque l'on utilise les

standard, tandis que le coefficient de régression correspondant est fixé à 0,1.

Ainsi qu'on peut le voir aux tableaux 3 et 4, l'estimateur entièrement pondéré avec mise à l'échelle parvient à éliminer le biais découlant de la nature informative du plan. La variance d'échantillonnage est plus élevée par rapport à l'estimateur non pondéré, en particulier dans le cas du coefficient de régression au niveau du sujet. Les résultats obtenus avec l'estimateur pondéré sont adéquats dans l'ensemble.

Tableau 3

Simulation – Moyennes et écarts types (entre parenthèses) des estimateurs ponctuels du coefficient de régression de la covariable au niveau du sujet (valeur réelle = 0,1; nombre de répétitions = 1 000)

d'échantillonnage	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	Taille fixe, $n_f = 38$	Taille prop., $n_f = 0,4N_f$	Taille fixe, $n_f = 9$	Taille prop., $n_f = 0,1N_f$
	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle				
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,101 (0,028)	0,117 (0,040)	0,098 (0,050)	0,098 (0,052)
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,099 (0,026)	0,117 (0,043)	0,098 (0,052)	0,098 (0,052)
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,119 (0,083)	0,116 (0,089)	0,100 (0,104)	0,098 (0,107)
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,099 (0,055)	0,116 (0,089)	0,100 (0,104)	0,098 (0,107)

Tableau 4

Simulation – Moyennes et écarts types (entre parenthèses) des estimateurs ponctuels du coefficient de régression de la covariable au niveau de la grappe (valeur réelle = 1 000)

d'échantillonnage	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	Taille fixe, $n_f = 38$	Taille prop., $n_f = 0,4N_f$	Taille fixe, $n_f = 9$	Taille prop., $n_f = 0,1N_f$
	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle				
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,096 (0,119)	0,117 (0,130)	0,102 (0,142)	0,096 (0,158)
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,102 (0,142)	0,106 (0,133)	0,106 (0,142)	0,095 (0,158)
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,116 (0,141)	0,116 (0,141)	0,105 (0,150)	0,105 (0,150)
Plan	Plan non informatif aux deux niveaux	Estimateur non pondéré	Estimateur pondéré	Estimateur entièrement pondéré avec mise à l'échelle	0,094 (0,117)	0,106 (0,133)	0,106 (0,142)	0,094 (0,158)

4.2.5 Considérations générales

Nos simulations montrent que la méthode du PMV représente dans la plupart des cas une stratégie simple et efficace à l'égard des plans d'échantillonnage informatifs. La seule condition consiste à connaître les probabilités d'inclusion à chaque étape du processus d'échantillonnage (sauf lorsque le caractère informatif du plan ne touche pas tous les niveaux).

Dans nos simulations, l'estimateur entièrement pondéré ayant fait l'objet d'une mise à l'échelle a donné de bons résultats relativement aux paramètres de régression, produisant un biais faible et entraînant une hausse modérée

de la variance d'échantillonnage (dans certains cas, la variance a même diminué). Même lorsque la pondération est inopérante, la perte d'efficacité attribuable à l'incorporation de coefficients de pondération mis à l'échelle est très faible.

Bien que la pondération semble être toujours efficace lorsque l'on fait l'estimation des paramètres de régression, il faut porter attention à la taille de l'échantillon dans le cas de la composante de la variance ω_0 : en effet, la pondération donne des résultats adéquats uniquement lorsque la grappe est de grande taille, permettant ainsi une bonne représentation de la structure de la variance complexe. Mais la taille de l'échantillon est également un facteur crucial pour l'estimation de ω_0 quand toutes les hypothèses de base du modèle ordinal multiniveaux sont vérifiées.

Les écarts rattachés à la catégorie de grappe dans l'échantillon – à taille fixe ou variable – sont minimes, les grappes de taille égale permettant d'obtenir des estimations légèrement meilleures; cela dit, ainsi que nous l'avons déjà indiqué, les écarts importants sont en grande partie attribuables à la taille moyenne des grappes composant l'échantillon.

Les résultats de notre étude de simulation confirment les observations de Pfeffermann et coll. (1998) au sujet du modèle linéaire à l'ordonnée à l'origine aléatoire : les estimateurs à pondération probabiliste sont efficaces pour l'ordonnée à l'origine, tandis qu'un biais significatif demeure lorsque l'on fait l'estimation des composantes de la variance pour un échantillon de petite taille. Ainsi que l'on pouvait s'y attendre, lorsque l'on passe d'un modèle linéaire à un modèle non linéaire, les résultats des estimateurs se dégradent légèrement, quoique l'importance et l'orientation du biais soient similaires dans les différents cas étudiés. De plus, les avantages de la mise à l'échelle sont confirmés.

Le degré d'augmentation de la variance d'échantillonnage attribuable à l'inclusion des coefficients de pondération se compare souvent aux résultats de Pfeffermann et coll. (1998), quoique nous ayons constaté parfois une réduction de la variance d'échantillonnage, entre autres dans le cas de l'ordonnée à l'origine, lorsque les coefficients de pondération sont mis à l'échelle et que le plan est informatif aux deux niveaux. L'une des différences intéressantes par rapport à l'étude de Pfeffermann et coll. (1998) a trait au rôle de la mise à l'échelle dans la réduction de la variance d'échantillonnage : la mise à l'échelle semble plus efficace dans le cas du modèle binaire que dans celui du modèle linéaire.

Nous l'avons dit, l'élément crucial lorsqu'il est question du modèle binaire à l'origine aléatoire est la question de l'estimation de la variance au niveau de la grappe ω_0 , qui demeure difficile lorsque le plan n'est pas informatif. À partir de la formulation des seuls exposés à la section 2, ω_0 est défini comme étant ω/σ , de sorte que l'estimation de

Tableau 1
Simulation – Moyennes et écarts types (entre parenthèses) des estimateurs ponctuels de l'ordonnée à l'origine
(valeur réelle = 0; nombre de répétitions : 1 000)

Plan d'échantillonnage		Estimateur non pondéré	Estimateur pondéré au niveau de la grappe	Estimateur entièrement pondéré sans mise à l'échelle	Estimateur entièrement pondéré avec mise à l'échelle
Plan informatif aux deux niveaux					
Taille fixe, $n_j = 38$	-0,120 (0,212)	-0,411 (0,202)	0,014 (0,193)	0,015 (0,188)	0,015 (0,188)
Taille proportionnelle, $n_j = 0,4N_j$	-0,163 (0,212)	-0,453 (0,200)	0,018 (0,190)	0,021 (0,183)	0,021 (0,183)
Taille fixe, $n_j = 9$	-0,214 (0,204)	-0,512 (0,190)	-0,062 (0,258)	0,000 (0,185)	0,000 (0,185)
Taille proportionnelle, $n_j = 0,1N_j$	-0,164 (0,220)	-0,450 (0,209)	-0,074 (0,294)	0,008 (0,203)	0,008 (0,203)
(niveau 2)					
Taille fixe, $n_j = 38$	0,281 (0,169)	0,018 (0,168)	0,017 (0,170)	0,017 (0,169)	0,017 (0,169)
Taille proportionnelle, $n_j = 0,4N_j$	0,274 (0,169)	0,014 (0,178)	0,014 (0,182)	0,014 (0,181)	0,014 (0,181)
Taille fixe, $n_j = 9$	0,274 (0,187)	0,010 (0,195)	0,010 (0,212)	0,009 (0,196)	0,009 (0,196)
Taille proportionnelle, $n_j = 0,1N_j$	0,269 (0,179)	0,007 (0,179)	0,007 (0,203)	0,006 (0,182)	0,006 (0,182)
Plan non informatif					
Taille fixe, $n_j = 38$	0,000 (0,108)	0,000 (0,114)	0,001 (0,115)	0,001 (0,115)	0,001 (0,115)
Taille proportionnelle, $n_j = 0,4N_j$	0,003 (0,113)	0,004 (0,120)	0,003 (0,123)	0,003 (0,122)	0,003 (0,122)
Taille fixe, $n_j = 9$	-0,007 (0,108)	-0,009 (0,115)	-0,010 (0,125)	-0,010 (0,117)	-0,010 (0,117)
Taille proportionnelle, $n_j = 0,1N_j$	-0,002 (0,110)	-0,002 (0,114)	-0,004 (0,132)	-0,003 (0,117)	-0,003 (0,117)

Tableau 2
Simulation – Moyennes et écarts types (entre parenthèses) des estimateurs ponctuels de l'écart type de deuxième niveau
(valeur réelle = 0,632; nombre de répétitions : 1 000)

Plan d'échantillonnage		Estimateur non pondéré	Estimateur pondéré au niveau de la grappe	Estimateur entièrement pondéré sans mise à l'échelle	Estimateur entièrement pondéré avec mise à l'échelle
------------------------	--	------------------------	---	--	--

Plan informatif aux deux niveaux					
Taille fixe, $n_j = 38$	0,671 (0,106)	0,638 (0,112)	0,637 (0,137)	0,604 (0,128)	0,604 (0,128)
Taille proportionnelle, $n_j = 0,4N_j$	0,673 (0,108)	0,636 (0,112)	0,645 (0,142)	0,592 (0,130)	0,592 (0,130)
Taille fixe, $n_j = 9$	0,644 (0,145)	0,584 (0,172)	0,920 (0,289)	0,536 (0,222)	0,536 (0,222)
Taille proportionnelle, $n_j = 0,1N_j$	0,598 (0,164)	0,546 (0,183)	1,002 (0,317)	0,498 (0,242)	0,498 (0,242)
Plan informatif uniquement au niveau de la grappe (niveau 2)					
Taille fixe, $n_j = 38$	0,595 (0,100)	0,596 (0,110)	0,605 (0,111)	0,601 (0,111)	0,601 (0,111)
Taille proportionnelle, $n_j = 0,4N_j$	0,582 (0,096)	0,582 (0,115)	0,603 (0,113)	0,596 (0,113)	0,596 (0,113)
Taille fixe, $n_j = 9$	0,547 (0,121)	0,548 (0,135)	0,671 (0,144)	0,563 (0,133)	0,563 (0,133)
Taille proportionnelle, $n_j = 0,1N_j$	0,538 (0,122)	0,535 (0,142)	0,696 (0,158)	0,551 (0,139)	0,551 (0,139)
Plan non informatif					
Taille fixe, $n_j = 38$	0,611 (0,086)	0,612 (0,092)	0,621 (0,090)	0,617 (0,091)	0,617 (0,091)
Taille proportionnelle, $n_j = 0,4N_j$	0,609 (0,084)	0,606 (0,088)	0,626 (0,088)	0,618 (0,088)	0,618 (0,088)
Taille fixe, $n_j = 9$	0,561 (0,105)	0,561 (0,112)	0,685 (0,119)	0,575 (0,111)	0,575 (0,111)
Taille proportionnelle, $n_j = 0,1N_j$	0,551 (0,109)	0,546 (0,113)	0,703 (0,134)	0,559 (0,112)	0,559 (0,112)

4.2.4 Simulations additionnelles à l'aide du modèle

avec des covariables

Des simulations additionnelles ont été effectuées pour évaluer les résultats de l'estimateur entièrement pondéré avec mise à l'échelle dans le modèle avec une covariable au niveau de la grappe et une au niveau du sujet. Le modèle est principal, exception faite de la covariable incluse à chaque niveau hiérarchique. Les valeurs de chacune des covariables sont produites à partir d'une distribution gaussienne

Nos résultats sont exposés et commentés en fonction des trois scénarios suivants : 1) *scénario de base* : le plan d'échantillonnage n'est pas informatif. Dans cette situation,

les hypothèses de base qui sous-tendent le modèle binaire,

donc consentent servir de point de référence pour évaluer

les résultats subséquents. 2) *Scénario du plan informatif*

sans pondération : le plan d'échantillonnage est informatif,

et l'estimateur n'est pas pondéré. Dans cette situation, les

hypothèses de base qui sous-tendent le modèle binaire à

l'ordonnée à l'origine aléatoire sont infirmées en raison du

caractère informatif du plan, et aucun ajustement n'est

effectué. 3) *Scénario du plan informatif avec pondération* :

Le plan d'échantillonnage est informatif et l'estimateur est

pondéré. Ici encore, les hypothèses de base sous-jacentes au

modèle binaire à l'ordonnée à l'origine aléatoire sont

infirmées, mais les coefficients de pondération sont

appliqués à titre d'ajustement pour contrebalancer le biais

existant au niveau de l'estimateur standard.

4.2.1 Scénario de base

Lorsque le plan d'échantillonnage n'est pas informatif, l'estimateur non pondéré du maximum de vraisemblance standard est asymptotiquement non biaisé (tableaux 1 et 2, lignes 9-12, colonne 1). Toutefois, dans le cas de petits échantillons ($n_j = 9$ et $n_j 0,1N_j$), l'estimation de ω_0 est assortie d'un biais négatif assez prononcé.

Si les coefficients de pondération sont appliqués alors qu'il n'y a pas lieu de procéder à un ajustement au titre de l'effet du plan (tableaux 1 et 2, lignes 9-12, colonnes 2 à 4), on observera une légère augmentation de la variabilité des estimateurs, plus prononcée lorsque l'estimateur entièrement pondéré sans mise à l'échelle est utilisé pour de petits échantillons. On remarquera que, toujours dans le cas de petits échantillons, l'estimateur entièrement pondéré sans mise à l'échelle de ω_0 est biaisé par excès.

4.2.2 Scénario du plan informatif sans pondération

Le caractère informatif du plan d'échantillonnage se traduit par des estimations biaisées et instables. Le biais demeure observable dans le cas de grands échantillons (tableaux 1 et 2, lignes 1 à 8, colonne 1). Les conclusions demeurent les mêmes pour les deux types de plan informatif, quoique le biais ait tendance à être de signe différent. En outre, le caractère informatif du plan fait augmenter la variabilité de l'estimateur standard par rapport au scénario de base : notamment, lorsque le plan est informatif aux deux niveaux, l'erreur type de l'estimateur de β_0 double.

4.2.3 Scénario du plan informatif avec pondération

Estimation de β_0

Les résultats présentés au tableau 1 montrent que, lorsque le plan est informatif, l'ajustement par pondération

constitue une méthode efficace pour éliminer le biais de l'estimation de β_0 . Notamment, lorsque le plan est informatif uniquement au niveau 2 (tableau 1, lignes 5-8, colonnes 2 à 4) et que les coefficients de pondération sont appliqués uniquement à ce niveau (estimateur pondéré au niveau de la grappe), le biais de l'estimation est corrigé sans qu'il y ait une hausse importante de la variance d'échantillonnage. Également, le résultat est valide dans le cas d'estimateurs entièrement pondérés (avec ou sans mise à l'échelle), et le biais est aussi corrigé dans le cas de petits échantillons.

Lorsque le plan est informatif aux deux niveaux (tableau 1, lignes 1-4, colonnes 2 à 4) et que les coefficients de pondération sont aussi appliqués aux deux niveaux (estimateurs entièrement pondérés), le biais au niveau de l'estimation de β_0 est corrigé. En outre, les estimateurs entièrement pondérés présentent une variance d'échantillonnage moins marquée que ceux qui ne sont pas pondérés, exception faite des estimateurs non mis à l'échelle dans le cas des petits échantillons. La mise à l'échelle est préférable particulièrement pour les petits échantillons, car cela permet d'obtenir un estimateur non biaisé et une variance d'échantillonnage beaucoup plus basse. Il faut mentionner que, lorsque le plan est informatif aux deux niveaux, l'estimateur pondéré au niveau de la grappe donne de moins bons résultats que l'estimateur standard non pondéré.

Estimation de ω_0

Les résultats présentés au tableau 2 à propos de ω_0 , sont passablement plus difficiles à interpréter (tableau 2, lignes 1-8, colonnes 2 à 4). Il faut remarquer d'abord que, dans le scénario de base, l'estimation de ω_0 est biaisée, surtout dans le cas de petits échantillons. Par conséquent, on considérera que l'ajustement par pondération est efficace s'il permet de reproduire le biais observé dans le scénario de base. Selon ce principe, le comportement de l'estimateur entièrement pondéré avec mise à l'échelle s'avère adéquat dans à peu près toutes les situations, exception faite des petits échantillons où le plan est informatif aux deux niveaux. Dans ce dernier cas, il faut aussi tenir compte d'un nombre non négligeable de répétitions qui ont donné une estimation zéro pour ω_0 (4,5 % pour le plan avec taille fixe et 2 % pour celui avec taille proportionnelle). Le problème des estimations nulles ne se pose pas avec l'estimateur entièrement pondéré sans mise à l'échelle; toutefois, exception faite d'une variance plus importante que dans le cas de l'estimateur avec mise à l'échelle, cet estimateur a tendance à surestimer ω_0 , indiquant un biais relatif de quelque 50 % pour les petits échantillons lorsque le plan est informatif aux deux niveaux. Précisons également que l'estimateur entièrement pondéré avec mise à l'échelle donne de meilleurs résultats que l'estimateur pondéré au niveau de la grappe, et ce, même lorsque le plan est informatif uniquement au niveau 2.

simples de tailles $0,25n_j$ et $0,75n_j$ ont été sélectionnés dans les strates respectives. Les tailles n_j étaient fixes, $n_j = n_0$ ou proportionnelles à N_j .

b) *Plan informatif au niveau 2* : le plan est le même que celui décrit précédemment, sauf que l'échantillonnage aléatoire simple a été utilisé pour la sélection des unités du niveau 1 à l'intérieur de chaque grappe échantillonnée.

c) *Plan non informatif* : le plan est le même que celui décrit au point b), sauf que la taille X_j est égale à N_j .

L'étude de simulation portait sur des échantillons où m est égal à 35 grappes, le nombre d'unités élémentaires étant variable : les gros échantillons avaient une taille fixe $n_j = n_0 = 38$ et une répartition proportionnelle $n_j = 0,4N_j$, les petits échantillons avaient une taille fixe $n_j = n_0 = 9$ et une répartition proportionnelle $n_j = 0,1N_j$ (moyenne égale à 9 environ).

L'étude de simulation a été entièrement effectuée à l'intérieur du SAS (SAS Institute 1999), dont le macro-langage a servi à rédiger le code requis. L'ajustement des modèles a été fait à l'aide de la procédure NLMIXED (se reporter à l'annexe A), comprenant une version adaptée de la quadrature gaussienne à dix points et un quasi-algorithme dual de Newton, une convergence a été atteinte après un petit nombre d'itérations. Ainsi que cela est expliqué à l'annexe A, dans le but d'éviter les erreurs d'arrondissement grossières, les coefficients de pondération du niveau 2 ont été multipliés au préalable par un facteur $k = 10\,000$, puis la matrice de covariance estimative a été multipliée par ce même facteur.

4.2 Résultats

Les résultats des simulations sont présentés aux tableaux 1 et 2. Pour chaque plan d'échantillonnage, le comportement des estimateurs ponctuels de l'ordonnée à l'origine β_0 et celui de l'écart type de deuxième niveau ω_0 sont déterminés à l'aide de la moyenne et de l'écart type de leur distribution d'échantillonnage de Monte Carlo. Les estimateurs ponctuels étudiés sont l'estimateur du maximum de vraisemblance standard non pondéré ainsi que les trois versions pondérées suivantes de ce dernier : les coefficients de pondération sont appliqués uniquement au deuxième niveau (w_j variables et w_{ij} constantes); *b) estimateur entièrement pondéré sans mise à l'échelle* : les coefficients de pondération sont appliqués aux deux niveaux, et ceux du premier niveau sont mis à l'échelle; *c) estimateur entièrement pondéré avec mise à l'échelle* : les coefficients de pondération sont appliqués aux deux niveaux, et ceux du premier niveau sont mis à l'échelle conformément à l'équation (10), soit la deuxième méthode de mise à l'échelle dont traite Pfeffermann et coll. (1998).

1 000 fois pour chaque combinaison de taille d'échantillon et de type de plan informatif. Pour qu'il soit possible de comparer nos résultats avec ceux obtenus dans le cas du modèle linéaire multiniveaux, l'expérience a été élaborée à partir de l'exemple fourni par Pfeffermann et coll. (1998, section 7).

La présente étude de simulation a porté d'abord sur un cas simple du modèle défini à la section 2, c'est-à-dire le modèle probit binaire à l'ordonnée 1 l'origine aléatoire. Ce modèle compte uniquement deux catégories relativement à une variable réponse (c'est-à-dire $K=2$) et une erreur aléatoire gaussienne au niveau de la grappe. De façon à établir un parallèle avec l'étude de Pfeffermann et coll. (1998), le modèle visé par le plan de simulation principal se rapporte au modèle sans covariables; cependant, des simulations additionnelles sont effectuées dans le but d'évaluer les résultats des estimateurs prévus dans le modèle avec une covariable au niveau de la grappe et une autre au niveau du sujet.

Les valeurs de la variable réponse binaire X_{ij} ont été produites à l'aide du plan à deux degrés suivant, qui est similaire à celui de Pfeffermann et coll. (1998) :

— Au premier degré, pour obtenir les valeurs de la population finie X_j ($j=1, \dots, M; i=1, \dots, N_j$), on a produit d'abord une valeur à partir du modèle latent de superpopulation $X_{ij}^* = \beta + u_j + \varepsilon_{ij}$, où $u_j \sim N(0, \omega^2)$ et $\varepsilon_{ij} \sim N(0, \sigma^2)$, puis $X_{ij}^* = 0$ si $X_{ij}^* \leq 0$ ou $X_{ij}^* = 1$ si $X_{ij}^* > 0$ (appelons que le modèle binaire comporte un seul sujet, qui est établi à zéro à des fins d'identification). Les valeurs des paramètres du modèle latent qui ont été employées dans la simulation sont $\beta = 0$, $\omega^2 = 0,2$ et $\sigma^2 = 0,5$, de sorte que les paramètres estimables à partir du modèle binaire sont $\beta_0 = \beta/\sigma = 0$ et $\omega_0 = \omega/\sigma = 0,632$ (se reporter à l'expression (3)). La structure hiérarchique de la population comprend $M = 300$ grappes, tandis que les tailles N_j des grappes ont été déterminées par $N_j = 75 \exp(u_j)$, où u_j est obtenu à partir de $N(0, \omega^2)$, tronqué vers le bas par $-1,5\omega$ et vers le haut par $1,5\omega$. Notre population N_j se situe donc à l'intérieur de la fourchette [38, 147], la moyenne se situant à 80 environ.

— Au deuxième degré, une fois obtenues les valeurs de la population finie, nous avons adopté l'un des plans d'échantillonnage suivants :

a) *Plan informatif aux deux niveaux* : on a d'abord choisi m grappes assorties de probabilités proportionnelles à une mesure de la taille X_j , c'est-à-dire $\pi_j = mX_j / \sum_{j=1}^M X_j$, la mesure X_j a été établie de la même manière que N_j , sauf que u_j est remplacé par n_j , soit l'effet aléatoire au niveau 2. Les unités X_j sont l'effet aléatoire au niveau 2. Les unités élémentaires de la j grappe échantillonnée ont ensuite été subdivisées en deux strates selon que $\varepsilon_{ij} > 0$ ou que $\varepsilon_{ij} \leq 0$, et des échantillons aléatoires

constitue un ensemble d'échantillons bootstrap indépendants à partir de la population finie artificielle et, pour chaque échantillon bootstrap, une estimation du paramètre cible est calculée; c) on obtient l'estimation de la variance du bootstrap, qui correspond à la variance de la distribution observée des estimations bootstraps.

On peut produire la population finie artificielle de la manière suivante : i) dans la j^{e} grappe échantillonnée, n_j est reproduite $w_j^{1/j}$ fois, le coefficient de pondération étant arrondi à l'entier le plus proche, ce qui donne une grappe artificielle composée de quelque N_j unités élémentaires; ii) chacune des m grappes artificielles est reproduite w_j fois, les coefficients de pondération étant arrondis à l'entier le plus proche, ce qui donne une population artificielle de quelque M grappes. Les échantillons sont ensuite sélectionnés à partir de la population artificielle, selon la procédure suivante : i) m grappes sont échantillonnées, les probabilités étant proportionnelles à π_j ; ii) dans la j^{e} grappe échantillonnée, n_j unités élémentaires sont échantillonnées, les probabilités étant proportionnelles à $\pi_j^{1/j}$.

Lorsque la fraction d'échantillonnage m/M est faible, la plus grande partie de la variance sera attribuable à l'échantillonnage des grappes, de sorte que la procédure bootstrap décrite précédemment pourra être simplifiée par la suppression des étapes relatives aux unités élémentaires – c'est-à-dire l'étape ii) de l'élaboration de la population artificielle et l'étape ii) du processus de rééchantillonnage.

L'estimateur de variance jackknife, envisagé par Korn et Graubard (2003), est une technique de rééchantillonnage plus simple aux fins d'estimation de la variance. Dans le cas de plans avec grappes, la technique utilisée comprend le calcul de la variance à partir d'un ensemble d'estimations ponctuelles obtenues en supprimant une grappe à la fois, et ce, malgré le fait que les résultats de la technique jackknife avec des données corrélées ne sont pas toujours satisfaisants (Shao et Tu 1995). Dans notre étude de simulation, l'estimateur de variance jackknife semble peu fiable, de sorte que nous ne l'avons pas utilisé. Il faudra procéder à des recherches additionnelles pour évaluer l'essai de possibilités de la technique jackknife et faire l'essai de certaines modifications pertinentes de cette technique.

4. ETUDE DE SIMULATION

4.1 Plan de l'expérience

L'expérience est conçue en fonction du plan à deux degrés dont on a fait l'hypothèse relativement aux variables observées : d'abord, les valeurs de la population finie sont produites à l'aide du modèle de superpopulation adéquat (premier degré); un échantillon informatif ou non informatif est ensuite sélectionné à partir de la population finie (deuxième degré), un seul échantillon étant constitué par population. Le plan de sélection à deux degrés a été répété

comme l'indiquent Korn et Graubard (2003). Nous reviendrons sur ce sujet à la section 4.

3.3 Technique d'estimation

La maximisation du logarithme du rapport de vraisemblance pondéré (7) passe par le calcul de plusieurs intégrales qui n'ont pas de solution analytique, de sorte qu'il faut employer une technique d'approximation numérique. Lorsque la dimensionnalité des intégrales est réduite, la quadrature gaussienne s'avère une technique simple et très précise; elle repose sur une sommation effectuée à partir d'un ensemble pertinent de points. La procédure NLMIXED du SAS (SAS Institute 1999) est une approche générale permettant d'ajuster les modèles non linéaires à effets aléatoires au moyen d'une version adaptative de la quadrature gaussienne. Différentes techniques d'optimisation peuvent être utilisées pour effectuer la maximisation; la technique par défaut, employée dans les simulations que nous commentons à la section 4, est le quasi-algorithme dual de Newton; l'adjectif dual signifie ici que la maximisation a trait au facteur de Cholesky d'un hessien approximatif (SAS Institute 1999).

La procédure NLMIXED ne comporte pas d'options relatives à une estimation du PMV, mais il est néanmoins possible d'incorporer des coefficients de pondération à la vraisemblance, en employant différentes astuces en ce qui touche les coefficients de pondération des niveaux 1 et 2, ainsi que cela est expliqué à l'annexe A.

3.4 Estimation de la variance

Dans le cadre classique du maximum de vraisemblance, on effectue l'estimation de la matrice de covariance des estimateurs en inversant la matrice d'information. Cet estimateur courant ne convient toutefois pas lorsque l'on utilise la méthode du PMV, car il ne tient pas compte de la variabilité découlant du plan d'échantillonnage. En vue d'obtenir une matrice de covariance plus fiable, Skinner (1989) a proposé le recours à un estimateur sandwich robuste, qui est également employé par Pfeiffermann et coll. (1998). Ainsi que nous l'avons mentionné à la section 3.3, la procédure NLMIXED du SAS permet d'ajuster le modèle de covariance ainsi estimé, obtenue par l'inversion de la matrice d'information, donnera alors probablement des résultats trompeurs lorsque l'on veut mesurer la variabilité réelle des estimateurs du PMV. Le calcul des estimateurs sandwich dans le cadre du SAS n'est pas une tâche facile. Il existe toutefois une solution à la fois simple et efficace – quoiqu'elle requière un certain travail de programmation – qui consiste à estimer la variance de façon empirique à l'aide de la technique bootstrap applicable aux populations finies (Sæmndal, Swensson et Wretman 1992). Cette technique comporte les étapes suivantes : a) à partir des données de l'échantillon, on élabore une population finie artificielle, censée reproduire la population réelle; b) on

Si les conditions ne sont pas sévères, la solution θ^{PML} de l'équation d'estimation $U(\theta) = 0$ sera convergente selon le plan pour l'estimateur du maximum de vraisemblance θ à l'égard d'une population finie, cet estimateur étant lui-même convergent selon le modèle par rapport au paramètre de superpopulation θ , de sorte que θ^{PML} constitue un estimateur convergent de θ pour ce qui est de la distribution mixte relativement au modèle et au plan (Pfeffermann 1993). On peut aussi obtenir des estimateurs généraux à pondération probabiliste à l'égard de modèles multiniveaux non linéaires en pondérant des fonctions d'estimation pertinentes, comme l'ont fait Singh, Folsom et Vaish (2002) dans le contexte de l'estimation régionale.

Pour utiliser la méthode du PMV, il faut connaître les probabilités d'inclusion aux deux niveaux. L'utilisation de coefficients de pondération au premier niveau seulement ou au second niveau seulement risque de s'avérer insuffisante et pourrait même empirer les choses, ainsi que nous avons pu le constater lors de nos simulations.

3.2 Mise à l'échelle des coefficients de pondération

L'un des points abordés par Pfeffermann et coll. (1998) ainsi que par Korn et Graubard (2003) et prêtant à controverse est la mise à l'échelle des coefficients de pondération en vue d'obtenir des estimateurs présentant un biais peu ou nul. On remarquera que le fait d'incorporer des coefficients de pondération au logarithme du rapport de vraisemblance suppose que l'on utilise un estimateur de la fonction de pondération de la population qui est cohérent avec le plan. Il est d'ailleurs possible de formuler la fonction de pondération de la population $U(\theta) \equiv \partial/\partial \theta \log L(\theta)$ de la façon suivante :

$$\log L(\theta) = \int_{+\infty}^f w_j \log \int_{+\infty}^f \exp \left\{ \sum_{i=1}^I w_{ij} \log L_{ij}(\theta) | n \right\} \phi(n) dn, \quad (7)$$

où \sum^I désigne une somme relative à des unités d'échantillonage. On remarquera que le fait d'incorporer des coefficients de pondération au logarithme du rapport de vraisemblance suppose que l'on utilise un estimateur de la fonction de pondération de la population qui est cohérent avec le plan. Il est d'ailleurs possible de formuler la fonction de pondération de la population $U(\theta) \equiv \partial/\partial \theta \log L(\theta)$ de la façon suivante :

La méthode du pseudo-maximum de vraisemblance (PMV) (Skinner 1989) est une solution employée couramment pour résoudre ce problème. Son utilisation est toutefois plus difficile dans le cas de modèles multiniveaux, du fait que le logarithme du rapport de vraisemblance associé à la population ne correspond pas à la somme de l'apport des unités élémentaires mais est plutôt une fonction des sommes des unités des niveaux 1 et 2, ce que l'on peut voir en formulant le logarithme du rapport de vraisemblance (5) de la manière suivante :

$$\log L(\theta) = \sum_{j=1}^f \log \int_{+\infty}^f \exp \left\{ \sum_{i=1}^I \log L_{ij}(\theta) | n \right\} \phi(n) du. \quad (6)$$

où $L_{ij} = L_{ij}(\theta) | n$ dont l'estimateur $U(\theta)$ correspond selon la méthode d'Horvitz et Thompson :

$$\frac{\sum_{j=1}^f w_j \int_{+\infty}^f \exp \left\{ \sum_{i=1}^I w_{ij} \log L_{ij}(\theta) | n \right\} \phi(n) dn}{\int_{+\infty}^f w_j \log \int_{+\infty}^f \exp \left\{ \sum_{i=1}^I w_{ij} \log L_{ij}(\theta) | n \right\} \phi(n) dn} \quad (9)$$

qui correspond à la valeur obtenue en dérivant le logarithme du rapport de vraisemblance à pondération probabiliste (7).

$$w_{ij}^{\text{échelle}} = \frac{w_{ij}}{w_{i|j}}, \quad (10)$$

où $w_j = (\sum_{i=1}^I w_{ij})/n_j$, de sorte que, pour la j^{e} grappe, la somme des coefficients de pondération mis à l'échelle est égale à la taille de l'échantillon de grappes n_j . Nous n'entendons pas commenter ici les avantages relatifs des différents méthodes de mise à l'échelle, aussi limiterons-nous nos simulations aux coefficients de pondération mis à l'échelle (10), dont la signification peut être appréhendée de façon intuitive et qui ont donné de bons résultats dans l'étude de Pfeffermann et coll. (1998), quoiqu'ils produisent un biais important dans le cas de certains plans, dans l'étude de simulation dont il est question à la section 4, nous présentons les résultats du type suivant de mise à l'échelle (la deuxième méthode de mise à l'échelle dont traitent Pfeffermann et coll. 1998) :

Dans l'étude de simulation dont il est question à la section 4, nous présentons les résultats du type suivant de mise à l'échelle (la deuxième méthode de mise à l'échelle dont traitent Pfeffermann et coll. 1998) :

où $i = 1, 2, \dots, N_j$ unités élémentaires (sujets) pour la j^{e} grappe ($j = 1, 2, \dots, M$). Dans l'équation (1), \mathbf{x}_{ij} est un vecteur de covariable et β est le vecteur de pente correspondants; les variables aléatoires ε_{ij} et u_j représentent les perturbations au premier (sujet) et au deuxième (grappe) niveaux, respectivement, tandis que ω_j^2 est la composante

de la variance de deuxième niveau.

En ce qui touche les perturbations du modèle (1), nous

faisons les hypothèses standard suivantes : a) les variables ε_{ij} sont iid, avec moyenne zéro et variance inconnue σ^2 ; b) les variables u_j sont iid gaussiennes avec une moyenne zéro et une variance unitaire; c) les variables ε_{ij} et u_j sont mutuellement indépendantes.

On remarquera que le modèle (1) produit le cas le plus simple de modèle ordinal multinationaux, comportant seulement deux niveaux et un effet aléatoire unique sur l'ordonnée à l'origine; la progression vers les niveaux

sujets et vers de multiples effets aléatoires se fait en principe de façon directe (Gibbons et Hedeker 1997), mais la complexité des formules amène à examiner uniquement le cas le plus simple, qui est suffisant pour que l'on puisse étudier les principaux concepts.

La variable ordinale observée Y est associée à la variable latente X . Leur relation peut être formulée de la façon suivante :

$$\{Y_{ij} = k\} \Leftrightarrow \{\gamma_{k-1} < X_{ij} \leq \gamma_k\},$$

où les seuils remplissent la condition suivante : $-\infty = \gamma_0 < \gamma_1 < \dots < \gamma_{K-1} < \gamma_K = +\infty$. Par conséquent, la probabilité du modèle à l'égard du sujet i de la grappe j , subordonnée à u_j est :

$$P(X_{ij} = k | u_j) = P(\gamma_{k-1} < X_{ij} \leq \gamma_k | u_j)$$

$$= P(X_{ij} \leq \gamma_k | u_j) - P(X_{ij} \leq \gamma_{k-1} | u_j), \quad (2)$$

et

$$P(X_{ij} \leq \gamma_k | u_j) = P(\varepsilon_{ij} \leq \gamma_k - [\beta' \mathbf{x}_{ij} + \omega_j u_j])$$

$$= P\left(\frac{\varepsilon_{ij}}{\gamma_k} \leq \frac{1}{\gamma_k} [\beta' \mathbf{x}_{ij} + \omega_j u_j] - \frac{\omega_j}{\gamma_k}\right)$$

$$= F(\gamma_{\alpha, k} - [\beta_{\alpha}' \mathbf{x}_{ij} + \omega_{\alpha} u_j]), \quad (3)$$

où $F(\cdot)$ est la fonction de distribution du terme d'erreur normalisé de premier niveau ε_{ij}/σ . Tous les paramètres du modèle sont définis par rapport à l'inconnue σ , c'est-à-dire l'écart type du terme d'erreur de premier niveau, de sorte que seuls les ratios des paramètres du modèle à l'écart type nous utilisons la notation ψ_{α} pour indiquer que le paramètre ψ du modèle latent s'exprime en unités de σ .

3.1 Estimateurs du pseudo-maximum de vraisemblance (PMV)

3. ESTIMATION À PONDERATION PROBABILISTE

Faisons l'hypothèse que la population totale, composée de M grappes (unités de niveau 2) et de N_j unités élémentaires (sujets, ou unités de niveau 1) par grappe, n'est pas observée; on utilise plutôt le plan d'échantillonnage à deux degrés suivant :

- premier degré : m grappes sont sélectionnées, les probabilités d'inclusion étant π_j ($j = 1, \dots, M$);
- deuxième degré : n_j unités élémentaires sont sélectionnées à l'intérieur de la j^{e} grappe sélectionnée, les probabilités étant $\pi_{ij}^{(j)}$ ($i = 1, \dots, N_j$).

La probabilité d'inclusion inconditionnelle de l'échantillon est alors $\pi_{ij} = \pi_{ij}^{(j)} \pi_j$. Lorsque le processus d'échantillonnage est de nature informative, c'est-à-dire lorsque les éléments π_j ou $\pi_{ij}^{(j)}$ dépendent des perturbations du modèle, et par le fait même de la variable réponse, l'estimateur du maximum de

peuvent être estimés. Supposons maintenant que θ dénote le vecteur de tous les paramètres pouvant être estimés, y compris β_{α} , ω_{α} et les $K-2$ seuils $\{\gamma_{\alpha, k}; k = 2, \dots, K-1\}$ (on fixe la valeur de $\gamma_{\alpha, 1}$ à zéro pour rendre l'identification possible). La vraisemblance conditionnelle relativement au sujet i de la grappe j est :

$$L_{ij}(\theta | u_j) = \prod_{k=1}^K \left[P(X_{ij} = k | u_j) \right]^{d_{ik}}, \quad (4)$$

où $P(X_{ij} = k | u_j)$ est défini à l'aide des équations (2) et (3), tandis que d_{ik} est la fonction indicatrice de l'événement $\{X_{ij} = k\}$. Dès lors, la vraisemblance marginale à l'égard de la grappe j est :

$$L_j(\theta) = \int_{-\infty}^{\infty} \prod_{N_j} L_{ij}(\theta | u_j) \phi(u_j) du_j,$$

où ϕ est la fonction de densité gaussienne standard. Enfin, la vraisemblance marginale globale est :

$$L(\theta) = \prod_{j=1}^M L_j(\theta). \quad (5)$$

cas des modèles non linéaires, les conceptuels de MLWIN ont prévu une procédure de pondération qui correspond à celle utilisée pour les modèles linéaires, certaines solutions spéciales étant prévues pour la variation de niveau 1 : par exemple, dans le cas des réponses binaires, les coefficients de pondération au niveau du sujet sont incorporés au dénominateur binomial. La méthode proposée est simple à mettre en application, mais ses propriétés n'ont pas encore été examinées. En outre, Renard et Molienberghs (2002) mentionnent le cas d'une application où l'utilisation de l'algorithme en question pour effectuer la pondération dans le cadre de modèles binaires multiniveaux n'a pas donné lieu à une convergence ou a produit des résultats non plausibles.

L'étude de simulation que nous allons utiliser pour évaluer les résultats des estimateurs du PMV concordera étroitement avec l'étude de Pfeffermann et coll. (1998), étant donné que ceux-ci emploient une approche similaire pour le modèle linéaire, ce qui permet certaines comparaisons intéressantes. Toutefois, il faut se rappeler à ce propos que, contrairement au modèle linéaire à deux niveaux, où les deux composantes de la variance peuvent faire l'objet d'estimations séparées, seul le ratio des deux composantes de la variance peut être estimé dans un modèle binaire à deux niveaux, ainsi que nous le commentons plus loin.

Korn et Graubard (2003) ont publié récemment une étude portant sur l'estimation des composantes de la variance; cette étude avait comme point de départ le fait que, relativement à de petits échantillons, on observait un biais marqué associé à plusieurs estimateurs pondérés des composantes de la variance proposés à titre d'ajustements dans le cas de plans informatifs (Graubard et Korn 1996). Bien que son sujet soit le même, l'étude de ces auteurs diffère de la nôtre sous bien des aspects : a) tout comme Pfeffermann et coll. (1998), ils se penchent uniquement sur les modèles multiniveaux linéaires; b) relativement à ces modèles, ils étudient d'abord l'estimation non biaisée des composantes de la variance dans de petits échantillons (de fait, ils proposent certains estimateurs à l'égard des composantes de la variance et se contentent d'esquisser la manière d'obtenir des estimateurs similaires pour un modèle linéaire avec des covariables, sans en tester l'efficacité; de toute manière, leur application à des modèles multiniveaux non linéaires n'est pas une chose facile à accomplir); c) les principaux estimateurs proposés par Korn et Graubard (2003), qui sont examinés en forme analytique, ont donné de bons résultats même dans le cas d'échantillons de petite taille. Toutefois, ils reposent sur les probabilités d'inclusion conjointe par paires. Lorsque ces probabilités ne sont pas connues, ce qui est souvent le cas, les auteurs proposent une autre approche, qui produit un biais important quand le nombre de grappes échantillonnées est peu élevé (33 dans leur plan de simulation). Avec la méthode du PMV utilisée

dans nos travaux, il n'est pas nécessaire de recourir à des probabilités d'inclusion conjointe. d) Le plan informatif utilisé par Korn et Graubard (2003) pour leur étude de simulation diffère nettement du nôtre; en effet, dans leur approche, le sous-échantillonnage des unités dépend du fait que les erreurs aléatoires du modèle soient supérieures à un seuil donné en valeur absolue, alors que, dans notre plan, le critère repose sur la question de savoir si les erreurs aléatoires sont élevées ou faibles. De ce fait, il est difficile de comparer les résultats obtenus.

En raison de l'utilisation très fréquente de modèles multiniveaux non linéaires dans de nombreux domaines d'application, il devient très important de disposer d'une méthode d'estimation pondérée à la fois générale et fiable, efficace et facile à appliquer, idéalement dans le cadre d'un logiciel statistique standard. La présente étude a pour objet de faire progresser les choses en ce sens.

Il convient de souligner que la méthode du PMV que nous utilisons a une portée assez générale, ce qui permet de l'appliquer à un large éventail de modèles. Ici, l'accent est mis sur les modèles qui produisent des réponses ordinales et binaires, car ils sont très fréquents et peuvent être représentés sous la forme d'un modèle linéaire assorti d'un ensemble de seuils par rapport à la réponse latente (se reporter à la section 2), ce qui facilite les comparaisons avec les résultats obtenus dans le cas du modèle linéaire. Précisons toutefois que la description de la méthode du PMV est tout à fait générale, et que la technique d'estimation fondée sur la procédure NLMIXED du SAS (exposée à l'annexe A) est facile à généraliser.

Notre étude est structurée de la façon suivante : la section 2 contient les définitions de base applicables au modèle ordinal multiniveaux; la section 3 décrit la méthode générale du PMV et donne certaines précisions concernant l'ajustement du modèle à l'aide de la procédure NLMIXED du SAS; à la section 4, nous évaluons, à l'aide d'une étude de simulation, les propriétés des différents estimateurs reliés au modèle binaire à l'ordonnée à l'origine aléatoire; nos observations finales sont présentées à la section 5.

2. LE MODÈLE ORDINAL MULTINIVEAUX

De manière à faciliter la comparaison avec les résultats relatifs au modèle linéaire (Pfeffermann et coll. 1998; Korn et Graubard 2003), il convient de formuler le modèle ordinal sous la forme d'un modèle linéaire latent reposant sur un ensemble de seuils. Supposons que la variable réponse ordinale observée, soit la variable X , où $k = 1, 2, \dots, K$ niveaux, est générée au moyen d'un ensemble de seuils à partir d'une variable continue latente Y , confortablement à un modèle des composantes de la variance (Hedeker et Gibbons 1994) :

$$Y_{ij} = \beta' x_{ij} + \omega u_{ij} + \varepsilon_{ij} \quad (1)$$

Estimation pondérée dans le cadre de modèles multivariés ordinaux et binaires sous un plan d'échantillonnage informatif

LEONARDO GRILLI et MONICA PRATESI¹

RÉSUMÉ

Les modèles multivariés sont souvent ajustés en fonction des données d'enquête recueillies dans le cadre d'un plan d'échantillonnage complexe à plusieurs degrés. Toutefois, lorsqu'un tel plan est informatif – en ce sens que les probabilités d'inclusion, même si on les subordonne aux covariables, dépendent de la variable réponse –, les estimateurs standard du maximum de vraisemblance seront biaisés. Dans la présente étude, nous inspirant de l'approche du pseudo-maximum de vraisemblance (PMV) de Skinner (1989), nous proposons une procédure d'estimation à pondération probabiliste dans le cadre de modèles multivariés ordinaux et binaires, de façon à supprimer le biais découlant du caractère informatif du plan d'échantillonnage. On utilise la valeur inverse des probabilités d'inclusion à chaque degré d'échantillonnage pour pondérer la fonction logarithmique de vraisemblance; les estimateurs pondérés que l'on obtient ainsi sont testés au moyen d'une étude de simulation dans le cas simple d'un modèle binaire à l'ordonnée à l'origine aléatoire, avec et sans covariables. Les estimateurs de la variance sont obtenus au moyen d'une procédure bootstrap. Pour maximiser le rapport de vraisemblance pondéré du modèle, nous avons recouru à la procédure NLMIXED du SAS, qui repose elle-même sur une version adaptée de la quadrature gaussienne. Également, l'estimation bootstrap des variances est effectuée dans l'environnement du SAS.

MOTS CLÉS : Plan informatif; modèle multivarié ordinaire; échantillonnage à plusieurs degrés; pseudo-maximum de vraisemblance; pondération.

1. INTRODUCTION

Dans de nombreux domaines de recherche, on utilisera fréquemment des modèles multivariés pour des réponses ordinales, y compris des réponses binaires à titre de cas spéciaux, pour modéliser des populations en grappes hiérarchisées. De fait, tant dans les sciences humaines qu'en biologie, on a souvent recouru, pour classer la situation ou la réponse d'un sujet, à deux catégories ou à un ensemble de catégories ordonnées (échelle ordinaire ou graduée). Parallèlement, on observe les sujets dans le contexte de leur appartenance à des groupes (par exemple, écoles, entreprises, cliniques, régions géographiques). La structure hiérarchique de la population sert souvent aussi à concevoir des plans d'échantillonnage à plusieurs degrés, les probabilités de sélection étant différentes à certains degrés du processus d'échantillonnage, voire à tous. Lors de l'analyse multivariée des données d'enquête, souvent, on ne tiendra pas compte des plans d'échantillonnage complexes même si cela risque d'engendrer la violation des hypothèses de base des modèles multivariés. En fait, lorsque des plans d'échantillonnage complexes sont utilisés, il se pourrait que les probabilités de sélection des sujets ainsi que des grappes ordonnées aux covariables, de la variable réponse; autrement dit, le plan d'échantillonnage pourrait être informatif. Relativement aux données qui sont mises en grappes et obtenues au moyen de plans informatifs à plusieurs degrés,

les ajustements proposés des modèles multivariés avaient principalement trait aux variables réponses continues. Pensons notamment à Pfeffermann, Skinner, Holmes, Goldstein et Rasbash (1998), qui proposent une procédure de pondération probabiliste des unités de premier et de deuxième niveaux, ajustées de façon à rendre compte des effets d'un plan informatif sur l'estimation de modèles à deux degrés avec variable réponse continue. Cette méthode, dite du pseudo-maximum de vraisemblance (PMV), consiste à formuler une expression en forme analytique de la vraisemblance associée au recensement, à procéder à l'estimation de la fonction logarithmique de vraisemblance, puis à effectuer la maximisation numérique de la fonction estimée. Dans cette méthode, il faut utiliser les coefficients de pondération d'échantillonnage à l'égard des éléments et grappes échantillonnées à tous les niveaux. Les auteurs conçoivent également les estimateurs « sandwich » applicables relativement aux variances des estimateurs. Les travaux de Pfeffermann et coll. (1998) visent essentiellement à incorporer le principe du PMV à l'algorithme des moindres carrés généralisés itérés (MCGI) (Goldstein 1986), qui peut être utilisé avec les modèles multivariés linéaires. L'algorithme des MCGI à pondération probabiliste est offert dans le programme MLwin (Rasbash, Browne, Goldstein, Yang, Plewisi, Healy, Woodhouse et Draper 1999), largement diffusé et utilisé. Précisons toutefois que son utilisation avec des modèles non linéaires n'est pas une mince tâche. En effet, dans le

Techniques d'enquête, juin 2004

PFEFFERMANN, D., et SVERCHKOV, M. (2003a). Fitting generalized linear models under informative probability sampling. Dans *Analysis of survey Data*. (Eds. C. Skinner et R. Chambers). New York: John Wiley & Sons, Inc. 175-195.

PFEFFERMANN, D., et SVERCHKOV, M. (2003b). Small area estimation under informative sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association (à paraître).

ROYAL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SÄRNDAAL, C.E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

Tableau 3 Erreurs de prévision moyennes et écart-type des moyennes (entre parenthèses) pour les trois modèles de prévision

Taille de l'échantillon	Modèle d'échantillon	Modèle de population	Modèle du complément d'échantillon
232	329,0 (2,2)	10,3 (2,3)	4,3 (2,3)
1 145	375,0 (0,9)	37,7 (1,1)	2,4 (1,1)
2 250	387,5 (0,6)	85,8 (0,7)	0,9 (0,8)

BIBLIOGRAPHIE

BREWER, K.R.W. (1993). Ratio estimation and finite populations: some results deducible from the assumptions of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

BREWER, K.R.W. (1999). Le calage esthique dans le cas de l'échantillonnage avec probabilités inégales. *Techniques d'enquête*, 25, 231-239.

CHAMBERS, R.L., DORMAN, A. et SVERCHKOV, M. (2003). Nonparametric regression with complex survey data. Dans *Analysis of Survey Data*, (Eds. C. Skinner et R. Chambers). New York: John Wiley & Sons, Inc. 151-174.

FULLER, W. (2003). Statistical analysis from complex survey data. Tutorial presented at the International Statistical Institute meeting, Berlin, Germany. Les diapositives du tutorial sont disponibles dans <http://cssm.iasstat.edu/academic/staff/fuller.html>.

HANSEN, M.H., MADOW, W.G. et TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (avec discussion). *Journal of the American Statistical Association*, 78, 776-807.

KIM, D.H. (2002). Bayesian and empirical Bayesian analysis under informative sampling. *Sankhyā B*, 64, 267-288.

KORN, E.L., et GRAUBARD, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.

PATKAR, Z., HIDIROGLOU, M. et LAVALLÉE, P. (2000). The methodology of the Workplace and Employee Survey. *Proceedings of the Second International Conference on Establishment Surveys*, 17-21 juin 2000, Buffalo, New York, American Statistical Association. 223-232.

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

PFEFFERMANN, D., et KRIEGER, A.M. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*, 13, 123-142.

PFEFFERMANN, D., KRIEGER, A.M. et RINOTT, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

PFEFFERMANN, D., et SVERCHKOV, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā, Series B*, 61, 166-186.

8. CONCLUSIONS

Une conclusion claire se dégage du tableau 3 : l'utilisation du modèle de population ou du modèle tenant compte des unités dans l'échantillon pour la prévision des valeurs y des unités à l'extérieur de l'échantillon peut se traduire par des biais appréciables. On notera que le biais induit par l'utilisation du modèle de population augmente avec la fraction de l'échantillonnage, ce qui correspond à la discussion précédente selon laquelle les différences entre le modèle d'échantillon et le modèle du complément d'échantillon ne se manifeste qu'aux tailles d'échantillon relativement importantes (voir le commentaire 2).

Dans cet article, nous utilisons des répartitions de l'échantillon et du complément de l'échantillon pour développer des prédicteurs convergents par rapport au plan d'échantillonnage, afin de prévoir des totaux de population finie. Nous démontrons que des prédicteurs connus, d'utilisation courante, constituent en fait des cas particuliers de notre théorie. Les EQM des nouveaux prédicteurs sont estimées par la combinaison d'un algorithme d'échantillonnage inverse et d'une méthode de rééchantillonnage. Comme le démontre notre exposé théorique et l'étude empirique, les prédicteurs des totaux de population finie qui requièrent uniquement la prévision des valeurs de résultat pour les unités à l'extérieur de l'échantillon fonctionnent mieux que les prédicteurs d'utilisation courante, même dans un cadre basé sur le plan d'échantillonnage, à moins que les fractions d'échantillonnage ne soient très faibles. Nous démontrons que les estimateurs EQM fonctionnent bien en terme de biais et quand on les utilise pour calculer les intervalles de confiance pour les totaux de population. Par conséquent, nous recommandons très fortement de poursuivre les expériences avec ces types de prédicteur et d'estimation des EQM.

REMERCIEMENTS

Les auteurs désirent remercier le rédacteur en chef adjoint et les deux examinateurs pour leurs commentaires et suggestions.

Tableau 2.1
Couverture nominale et empirique en pourcentage des intervalles de confiance, $n = 232$

Groupe	Predicteur	1,0	2,5	5,0	10,0	90,0	95,0	97,5	99,0
Pas de valeur x	\hat{Y}_{H-T}	2,5	3,5	5,5	10,0	90,0	97,0	99,0	99,5
	\hat{Y}_{EI}	0,5	2,0	4,0	8,0	88,5	91,5	95,5	98,0
2	\hat{Y}_{Hak}	0,5	2,0	4,0	8,0	88,5	91,5	95,5	98,0
	\hat{Y}_1	0,0	0,0	1,5	6,5	86,0	90,5	92,5	97,5
Régression	$\hat{Y}_{2, Reg}$	0,0	0,0	2,0	7,0	85,0	90,5	93,5	98,0
polynômiale du	\hat{Y}_3	0,0	0,5	2,5	6,5	87,5	91,0	95,0	98,5
troisième degré	\hat{Y}_{GREG}	0,0	0,0	2,0	7,0	85,0	90,5	93,5	98,0
3	$\hat{Y}_{2, Reg}$	0,0	1,0	2,5	7,0	87,0	91,5	96,5	98,0
	\hat{Y}_3	0,0	1,0	2,5	7,0	86,0	91,5	96,5	98,0
Régression simple	\hat{Y}_{GREG}	0,0	1,0	2,5	7,0	86,5	91,5	97,0	98,0

Tableau 2.2
Couverture nominale et empirique en pourcentage des intervalles de confiance, $n = 1145$

Groupe	Predicteur	1	2,5	5,0	10,0	90,0	95,0	97,5	99,0
Pas de valeur x	\hat{Y}_{H-T}	4,0	7,0	9,0	13,5	95,5	98,0	98,5	99,5
	\hat{Y}_{EI}	3,0	5,0	8,0	12,5	92,5	95,5	99,5	100,0
2	\hat{Y}_{Hak}	3,5	5,0	9,5	12,5	92,5	96,0	99,5	100,0
	\hat{Y}_1	0,5	2,0	5,0	7,5	86,5	93,5	96,0	97,0
Régression	$\hat{Y}_{2, Reg}$	0,5	3,0	6,0	9,0	86,5	94,5	96,5	97,0
polynômiale du	\hat{Y}_3	0,5	2,0	6,0	9,5	88,0	94,0	97,0	98,0
troisième degré	\hat{Y}_{GREG}	0,5	3,0	5,0	9,0	86,5	94,0	96,5	98,0
3	$\hat{Y}_{2, Reg}$	0,5	3,0	6,0	11,0	90,0	93,0	97,0	99,5
	\hat{Y}_3	0,5	2,5	5,5	10,5	90,0	94,0	97,0	99,5
Régression simple	\hat{Y}_{GREG}	1,0	3,0	6,0	11,0	90,5	94,0	97,5	99,0

Tableau 2.3
Couverture nominale et empirique en pourcentage des intervalles de confiance, $n = 2250$

Groupe	Predicteur	1,0	2,5	5,0	10,0	90,0	95,0	97,5	99,0
Pas de valeur x	\hat{Y}_{H-T}	0,5	1,0	5,5	11,0	95,0	97,5	99,0	99,5
	\hat{Y}_{EI}	1,0	3,0	5,5	9,0	91,5	96,0	99,0	99,5
2	\hat{Y}_{Hak}	1,0	2,5	5,5	9,0	93,0	97,0	98,5	99,5
	\hat{Y}_1	0,5	2,0	5,0	9,0	91,0	94,5	96,5	97,5
Régression	$\hat{Y}_{2, Reg}$	0,5	2,5	6,5	10,5	90,5	94,5	96,5	98,0
polynômiale du	\hat{Y}_3	0,5	2,0	7,5	12,5	91,5	95,5	96,5	97,5
troisième degré	\hat{Y}_{GREG}	0,5	2,0	6,0	11,0	91,0	94,5	96,0	98,0
3	$\hat{Y}_{2, Reg}$	1,0	3,0	6,0	11,0	91,0	95,0	97,5	99,0
	\hat{Y}_3	1,0	2,0	6,0	12,0	90,0	95,0	97,5	98,0
Régression simple	\hat{Y}_{GREG}	0,0	1,5	5,0	11,5	91,5	95,0	97,5	99,0

Tableau 1.1

Biais, REQ  et racine carr e de la moyenne des estimateurs de l'EQ , n = 232

Groupe	Pr�dicateur	Biais (e.-l.)	REQ�	$\sqrt{EQ�}$
1	\hat{Y}_{H-T}	-4,5 (1,6)	365,1	355,0
Pas de valeur x	\hat{Y}_{EI}	1,5 (2,9)	91,1	89,8
2	\hat{Y}_1	4,4 (2,0)	64,0	63,0
R�gression	$\hat{Y}_{2, Reg}$	3,5 (2,0)	63,4	62,4
polynomiale du	\hat{Y}_3	-0,3 (2,1)	65,4	65,0
troisi�me degr�	\hat{Y}_{GREG}	3,4 (2,1)	63,6	62,6
3	$\hat{Y}_{2, Reg}$	-2,3 (2,2)	68,0	66,2
R�gression simple	\hat{Y}_3	-0,3 (2,2)	68,6	67,4
	\hat{Y}_{GREG}	-2,3 (2,2)	68,3	66,5
Total de la « population » vraie = 2 710,7				

Tableau 1.2

Biais, REQ  et racine carr e de la moyenne des estimateurs de l'EQ , n = 145

Groupe	Pr�dicateur	Biais (e.-l.)	REQ�	$\sqrt{EQ�}$
1	\hat{Y}_{H-T}	-9,1 (5,0)	157,1	156,1
Pas de valeur x	\hat{Y}_{EI}	0,0 (1,1)	35,2	34,9
	\hat{Y}_{Hajek}	-0,1 (1,3)	39,5	39,3
2	\hat{Y}_1	3,0 (0,9)	27,6	28,1
R�gression	$\hat{Y}_{2, Reg}$	2,0 (0,9)	27,4	27,3
polynomiale du	\hat{Y}_3	0,5 (0,9)	27,4	27,7
troisi�me degr�	\hat{Y}_{GREG}	1,7 (0,9)	27,8	27,8
3	$\hat{Y}_{2, Reg}$	0,0 (1,0)	28,3	28,7
R�gression simple	\hat{Y}_3	0,1 (1,0)	28,2	28,9
	\hat{Y}_{GREG}	0,0 (2,0)	29,1	29,6
Total de la « population » vraie = 2 710,7				

Tableau 1.3

Biais, REQ  et racine carr e de la moyenne des estimateurs de l'EQ , n = 2 250

Groupe	Pr�dicateur	Biais (e.-l.)	REQ�	$\sqrt{EQ�}$
1	\hat{Y}_{H-T}	1,3 (2,7)	82,7	80,4
Pas de valeur x	\hat{Y}_{EI}	-0,2 (0,6)	18,5	18,8
	\hat{Y}_{Hajek}	0,1 (0,7)	23,5	23,8
2	\hat{Y}_1	1,3 (0,5)	17,5	17,3
R�gression	$\hat{Y}_{2, Reg}$	0,6 (0,5)	16,9	16,3
polynomiale du	\hat{Y}_3	-0,3 (0,5)	17,1	16,5
troisi�me degr�	\hat{Y}_{GREG}	0,5 (0,5)	17,9	18,3
3	$\hat{Y}_{2, Reg}$	-0,3 (0,5)	17,3	16,8
R�gression simple	\hat{Y}_3	-0,3 (0,5)	17,7	17,3
	\hat{Y}_{GREG}	-0,2 (0,6)	18,8	18,3
Total de la « population » vraie = 2 710,7				

2- Les prédicteurs des groupes 2 et 3 qui utilisent les valeurs auxiliaires fonctionnent mieux que les prédicteurs du groupe 1, particulièrement pour les échantillons de petite taille. Les prédicteurs du groupe 2 qui emploient le modèle de régression polynomiale du troisième degré (7.1) fonctionnent mieux que les prédicteurs correspondants du groupe 3, qui emploient le modèle de régression simple comme modèle de travail, mais cette différence s'atténue à mesure que la taille de l'échantillon augmente.

3- Cette étude donne un résultat important, à savoir : les prédicteurs $\hat{Y}_{2, Reg}$ et $\hat{Y}_{3, EI}$ (et également \hat{Y}_3 pour les échantillons de grandes tailles), qui prévoient seulement les valeurs y pour les unités à l'extérieur de l'échantillon, fonctionnent mieux que les autres prédicteurs dans leurs groupes respectifs (voir également ci-dessous). Comme le résume la remarque 7, cela est particulièrement vrai dans le cas des échantillons de grande taille. On notera que les différences entre $\hat{Y}_{2, Reg}$ et l'estimateur GREG pour $n = 1\ 145$ et $n = 2\ 250$ sont plus faibles dans le modèle polynomiale (groupe 2) qu'avec le modèle de régression simple (groupe 3), ce qui s'explique par la relation étroite entre la variable d'étude et les variables auxiliaires dans le modèle polynomiale. Le prédicteur \hat{Y}_3 est moins stable que $\hat{Y}_{2, Reg}$ pour $n = 232$, mais pour les deux autres tailles d'échantillon, les deux prédicteurs offrent une efficacité similaire.

4- Le prédicteur $\hat{Y}_{2, Reg}$ fonctionne un peu mieux que le prédicteur \hat{Y}_1 qui dépend du modèle et qui emploie les espérances $E(w_i | x_i)$ pour l'ajustement des poids d'échantillonnage. Nous n'avons pas d'explications claires de ces résultats, car comme l'ont illustré Pfeffermann et Sverchkov (1999) à l'aide des mêmes données, l'ajustement des poids d'échantillon améliore l'estimation des coefficients de régression d'une manière très notable.

Examinons ensuite les estimateurs de l'EQM.

5- Les estimateurs de l'EQM développés à la section 6 fonctionnent très bien pour tous les prédicteurs et avec toutes les tailles d'échantillon. Pour la taille d'échantillon $n = 232$, on constate une sous-estimation systématique de la valeur REQM atteignant jusqu'à 3 %, qui s'explique par le fait que la pseudo-population dans ce cas-ci est moins variable que la population d'étude réelle (voir la remarque 9). Les estimateurs de l'EQM sont

$$epm = \sum_{j=1}^{1000} (\hat{y}_j - y_j) / (N - n) \Big/ 1000$$

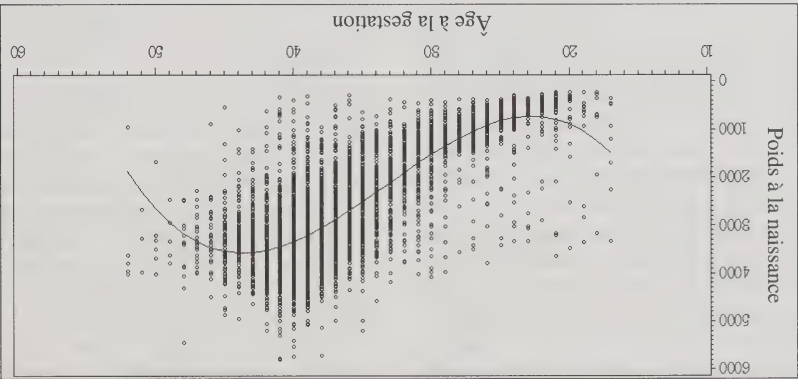
où S_j définit le $j^{ème}$ échantillon sélectionné. Les erreurs EPM sont illustrées pour les trois prédicteurs, tous utilisant le modèle de travail (7.1) et ayant la forme générale $\hat{y}_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3$, $j \neq s$. Pour le premier prédicteur, le vecteur $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ est estimé par la méthode des moindres carrés ordinaires, ce qui revient à utiliser le modèle d'échantillon; pour le deuxième prédicteur, β est estimé par l'estimateur B^{pw} , pondéré en probabilité, qui correspond à l'utilisation du modèle de population, tandis que le troisième prédicteur B^c qui est calculé de la même manière que B^{pw} , mais avec les poids $(w_i - 1)$, ce qui correspond à l'utilisation du modèle de complètement d'échantillon.

Afin de mieux illustrer ce point, nous présentons dans le tableau 3 l'erreur de prévision moyenne (epm) dans l'échelle originale (en grammes), pour les 1 000 échantillons quand on prédit les valeurs du complètement de l'échantillon; uniquement appréciables quand des fractions d'échantillonage ne sont pas négligeables.

Comme le sous-entend l'exposé théorique de notre article et comme l'illustre l'étude empirique, la prévision des valeurs y seulement pour les unités à l'extérieur de l'échantillon, à l'aide du modèle de complètement d'échantillon, donne des meilleurs prédicteurs pour le total de la population que si on cherche à prévoir toutes les valeurs de la population en utilisant un modèle de population, ce qui est implicitement fait quand on utilise des estimateurs GREG ou de Hajek. Il est manifeste que les différences sont beaucoup plus faibles pour les unités à l'extérieur de l'échantillon, à l'aide du modèle de complètement d'échantillon, que pour les unités à l'intérieur de l'échantillon, à l'aide du modèle de complètement d'échantillon. Les pourcentages empiriques sont quelque peu erratiques avec la taille d'échantillon $n = 232$, mais ils se stabilisent à mesure que la taille d'échantillon augmente, notamment avec l'utilisation des prédicteurs des deuxième et troisième groupes. Les pourcentages empiriques sont près des pourcentages nominaux, pour tous les prédicteurs, quand $n = 2\ 250$.

Une autre façon d'évaluer le biais des divers prédicteurs et vraies étant de nouveau de l'ordre de 3 %.

presque sans biais pour les autres tailles d'échantillon, la plus grande différence entre les valeurs REQM estimées et vraies étant de nouveau de l'ordre de 3 %.



$$\text{Mod le ajust  : } y_i = 17886 - 1827,7x_i + 61,2x_i^2 - 0,61x_i^3 + e_i$$

$$\text{Var}(e_i) = 603,2, \quad R^2 = 0,61$$

Figure 1. Nuage de points du poids   la naissance par rapport   l' ge   la gestation dans la « population » ( chantillon originale), et valeurs pr dites par r gression polynomiale du troisi me degr .

Les pr dicateurs envisag s dans cette  tude se divisent donc en trois groupes. Le premier comprend les pr dicateurs qui utilisent uniquement les valeurs y et les poids d' chantillonnage. Front part  de ce groupe l'estimateur Horvitz-Thompson $\hat{Y}_{H-T} = \sum_{i \in s} w_i y_i$, le pr dicateur \hat{Y}_{EI} d fini par (5.2) et l'estimateur de Hajek \hat{Y}_{Hajek} d fini par (5.3). Le deuxi me comprend les pr dicateurs qui utilisent le mod le de travail d fini par l' quation (7.1). Front part  de ce groupe les deux pr dicateurs de r gression \hat{Y}_1 et $\hat{Y}_{2, Reg}$ d finis par les  quations (4.4) et (4.8) respectivement, le pr dicateur avec correction de biais \hat{Y}_3 d fini par (4.12) et enfin l'estimateur GREG d fini par (4.14). Le troisi me groupe contient les m mes pr dicateurs que le deuxi me groupe (exception faite de \hat{Y}_1 , voir ci-dessus), mais ils sont bas s sur le mod le de r gression simple (c'est- -dire uniquement la premi re puissance de x).

$$EQM(\hat{Y}) = \frac{1}{1000} \sum_{r=1}^{1000} (\hat{Y}^{(r)} - Y)^2 \quad (7.2)$$

g n rique a  t  calcul  comme suit.

L' tude. Ainsi, l'E M « vraie » d'un pr dicateur \hat{Y} 1 000  chantillons s lectionn s   partir de la population   E M empiriques correspondantes bas es sur les l'efficacit  des estimateurs E M, nous avons calcul  les provenant de chaque pseudo-population. Afin d' valuer donc bas s sur seulement 200  chantillons bootstrap s lection al atoire de 200 des 1 000  chantillons, et sont estimateurs E M n'ont pu  tre calcul s que pour une limite de calcul machine dont nous disposons, les  tapes d crites   la section 6. Toutefois, en raison du temps  tude ont  t  estim s   l'aide de la proc dure en deux

Les E M de tous les pr dicateurs pr sent s dans cette  tude ont  t  estim s   l'aide de la proc dure en deux  tapes d crites   la section 6. Toutefois, en raison du temps limite de calcul machine dont nous disposons, les estimateurs E M n'ont pu  tre calcul s que pour une s lection al atoire de 200 des 1 000  chantillons, et sont donc bas s sur seulement 200  chantillons bootstrap provenant de chaque pseudo-population. Afin d' valuer l'efficacit  des estimateurs E M, nous avons calcul  les E M empiriques correspondantes bas es sur les 1 000  chantillons s lectionn s   partir de la population    tude. Ainsi, l'E M « vraie » d'un pr dicateur \hat{Y} g n rique a  t  calcul  comme suit.

1- Tous les pr dicateurs pr sent s dans cette  tude sont essentiellement sans biais selon le plan d' chantillonnage pour les trois tailles d' chantillon, peu importe le mod le de travail sous-jacent. Le pr dicateur \hat{Y}_1 pr sente un biais statistiquement significatif quand il est test    l'aide de la statistique t conventionnelle, mais le biais r el est n gligeable par rapport au total de la population vraie. (Le pr dicateur \hat{Y}_1 est le seul pr dicateur, dans la pr sente  tude, qui ne pr sente pas de convergence par rapport au plan d' chantillonnage).

Les trois commentaires suivants portent sur les E M des divers pr dicateurs.

Les principaux r sultats de cette  tude sont pr sent s dans les tableaux 1.1   1.3 (un tableau pour chaque taille d' chantillon). La troisi me colonne de chaque tableau indique pour chaque pr dicateur \hat{Y} le biais empirique, $[(\sum_{r=1}^R \hat{Y}^{(r)} / R) - Y]$, et l' cart-type (e.-t.) du biais empirique, qui  quivaut   $[\sum_{r=1}^R (\hat{Y}^{(r)} - Y)^2 / R^2]^{1/2}$; $\hat{Y}_R = \sum_{r=1}^R \hat{Y}^{(r)} / R$, $R = 1000$. Les deux colonnes suivantes indiquent respectivement la valeur RE M (racine carr e de l' quation 7.2) « vraie » (empirique), ainsi que la racine carr e de la moyenne des estimateurs bootstrap correspondants, d finis par (6.2). Nous pouvons tirer des tableaux 1.1   1.3 les principales conclusions suivantes :

7.2 R sultats de l' tude empirique

Les principaux r sultats de cette  tude sont pr sent s dans les tableaux 1.1   1.3 (un tableau pour chaque taille d' chantillon). La troisi me colonne de chaque tableau indique pour chaque pr dicateur \hat{Y} le biais empirique, $[(\sum_{r=1}^R \hat{Y}^{(r)} / R) - Y]$, et l' cart-type (e.-t.) du biais empirique, qui  quivaut   $[\sum_{r=1}^R (\hat{Y}^{(r)} - Y)^2 / R^2]^{1/2}$; $\hat{Y}_R = \sum_{r=1}^R \hat{Y}^{(r)} / R$, $R = 1000$. Les deux colonnes suivantes indiquent respectivement la valeur RE M (racine carr e de l' quation 7.2) « vraie » (empirique), ainsi que la racine carr e de la moyenne des estimateurs bootstrap correspondants, d finis par (6.2). Nous pouvons tirer des tableaux 1.1   1.3 les principales conclusions suivantes :

1- Tous les pr dicateurs pr sent s dans cette  tude sont essentiellement sans biais selon le plan d' chantillonnage pour les trois tailles d' chantillon, peu importe le mod le de travail sous-jacent. Le pr dicateur \hat{Y}_1 pr sente un biais statistiquement significatif quand il est test    l'aide de la statistique t conventionnelle, mais le biais r el est n gligeable par rapport au total de la population vraie. (Le pr dicateur \hat{Y}_1 est le seul pr dicateur, dans la pr sente  tude, qui ne pr sente pas de convergence par rapport au plan d' chantillonnage).

Les trois commentaires suivants portent sur les E M des divers pr dicateurs.

$$f^{pp}(y_i | \mathbf{x}_i) = \frac{E_s(Cw_i | y_i, \mathbf{x}_i) f_s(y_i | \mathbf{x}_i)}{E_s(Cw_i | \mathbf{x}_i)} = \frac{E_p(\pi_i | \mathbf{x}_i) f_s(y_i | \mathbf{x}_i)}{E_p(\pi_i | y_i, \mathbf{x}_i)} = f^p(y_i | \mathbf{x}_i), \quad (6.3)$$

Remarque 9. L'équation (6.3) porte uniquement sur la distribution marginale de $y_i | \mathbf{x}_i$. Tout comme avec la méthode bootstrap standard, l'application réussie de la procédure que nous proposons requiert que la taille de l'échantillon original soit suffisamment importante et que les mesures de l'échantillon soient approximativement indépendantes. Pfeffermann et coll. (1998) établissent les conditions dans lesquelles, pour des mesures de population indépendantes, les mesures d'échantillon sont « asymptotiquement indépendantes » dans des schémas d'échantillonnage couramment utilisés avec des probabilités de sélection intégrales.

Remarque 10. L'étape 1 est similaire et asymptotiquement équivalente au dédoublement des unités d'échantillon $i w_i$ fois. On notera toutefois que l'utilisation de cette procédure de duplication ne donne pas des pseudo-populations de taille N_i , à moins que $\sum_{i=1}^I w_i = N$. On ne voit pas clairement également comment établir la relation (6.3) quand on utilise cette procédure.

7. ILLUSTRATIONS EMPIRIQUES

7.1 Description de l'étude empirique

Afin d'illustrer l'efficacité des prédicteurs et des estimations des EQM connexes, traitées dans la section précédente, nous utilisons un ensemble de données réelles, recueillies dans le cadre de l'Enquête américaine nationale de 1988 sur la santé des mères et des nouveau-nés (*National Maternal Infant Health Survey*). L'enquête utilise un échantillon aléatoire stratifié disproportionné des actes de l'état civil, les strates étant définies par la *race de la mère* et le *poids de l'enfant à la naissance*; voir Korn et Graubard (1995) pour plus de détails. Pour l'étude empirique présentée dans cette section, nous considérons que les données d'échantillon constituent la « population » et nous avons sélectionné indépendamment 1 000 échantillons avec des probabilités proportionnelles à l'inverse des poids d'échantillonnage originaux, utilisant un schéma d'échantillonnage avec probabilité proportionnelle à la taille (PPT) systématique. La liste des « unités de population » a été tirée de manière aléatoire avant chaque sélection d'échantillon. Pour chaque échantillon, nous avons prélevé 10 000 dans l'étude présente) pour l'ensemble de la population, utilisant l'âge à la gestation comme variable auxiliaire (mesurée en

semaines). Les probabilités d'inclusion dans l'échantillon dépendent donc des valeurs de la variable étudiée qui définit les strates originales. On notera que bien que même si l'échantillon original est essentiellement un échantillon aléatoire stratifié, les poids d'échantillonnage varient en fait à l'intérieur même des strates, et c'est pourquoi nous avons utilisé l'échantillonnage PPT systématique pour l'étude de simulation. Nous avons considéré trois tailles d'échantillon différentes, soit $n = 232$, 1 145, 2 429. La taille de la « population » (échantillon original) est $N = 9\,948$. (Pour $n = 232$, $0,002 < \pi_i = \Pr(i \in s) < 0,15$. Pour $n = 1\,145$, $0,01 < \pi_i < 0,73$. Pour $n = 2\,429$, $0,03 < \pi_i < 0,99$ avec la moyenne $\bar{\pi} = 0,26$ et un écart-type $Std(\pi_i) = 0,29$. Dans ce dernier cas, certaines des unités ont été tirées avec une quasi-certitude).

Certains des prédicteurs étudiés dans le présent document (voir ci-dessous) requiert que l'on spécifie soit le modèle de l'échantillon, soit le modèle du complément de l'échantillon. Nous avons supposé pour les deux modèles une régression polynomiale du troisième degré,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \quad (7.1)$$

avec des valeurs résiduelles indépendantes et une variance constante. Pfeffermann et Sverdchikov (1999) ont trouvé que ce modèle représentait une bonne approximation des données de la « population » (échantillon d'origine) avec $R^2 = 0,61$ (voir la figure 1), et ils ont jugé qu'elle correspondait assez bien aux données de l'échantillon (avec des coefficients différents) pour plusieurs échantillons sélectionnés à partir de cette « population ». On notera par ailleurs que dans ce schéma d'échantillonnage fortement informatif, il est peu probable que le modèle d'échantillon, le modèle de population et le modèle du complément d'échantillon proviennent tous de la même famille, même si les paramètres sont différents. La présente étude permet donc d'étudier l'efficacité de divers prédicteurs quand un, plusieurs ou tous les trois modèles sont mal spécifiés. La question importante de la robustesse est de plus examinée par ajustement des modèles de régression simples, au lieu d'utiliser des régressions polynomiales du troisième degré, c'est-à-dire en omettant les puissances deux et trois de la variable auxiliaire. La seule exception est le prédicteur dépendant du modèle, \hat{Y}_i (équation 4.4), pour lequel aucun estimateur cohérent pour l'espérance $E_s(w_i | x_i)$ n'a pu être trouvé quand on le restreignait à une régression simple. (La méthode présentée dans Pfeffermann et Sverdchikov (1999) pour l'estimation de cette espérance suppose que les valeurs résiduelles du modèle de population sont normales. Cette hypothèse est valide quand on ajuste le modèle de régression polynomiale du troisième degré, mais elle est clairement violée quand on omet les puissances deux et trois de la variable auxiliaire).

et

$$E_c(\varepsilon_j | x_j) = E_p[1 - \pi(x_j) - K\varepsilon_j - K\varepsilon_j / (1 - \pi(x_j))] \varepsilon_j | x_j] \\ = -x_j^* / (1 - \pi(x_j)).$$

Il d coule donc que $E_{QM}^p(Y_{E, Reg} | D_s) = \sigma^2 \sum_{j \in s} x_j - \sum_{j \in s} [x_j^* / (1 - \pi(x_j))]$. Q.E.D.

Remarque 8: Pour l' chantillonnage *non informatif* et avec une valeur β connue, l'E M de pr vision pour le pr dicteur optimal $\hat{Y} = \sum_{i \in s} y_i + \beta \sum_{j \in s} x_j$ est $E_p[(\hat{Y} - Y)^2 | D_s] = \sigma^2 \sum_{j \in s} x_j$. Cette E M est plus grande que l'E M obtenue d'apr s le sch ma d' chantillonnage informatif d fini par le lemme, ce qui est  vident car ce dernier sch ma a tendance    chantillonner les unit s pr sentant les valeurs y les plus grandes, et donc les valeurs x les plus grandes, d'o  des  carts-types plus importants.

6. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE

Pour estimer $E_{QM}^p(Y | D_s) = E_p[(\hat{Y} - Y)^2 | D_s]$, pour les pr dicteurs \hat{Y}  tudi s   la section 4, il faut formuler des hypoth ses de mod le strictes qui peuvent  tre difficiles   valider. Cela est d  essentiellement au conditionnement   l'information sur le plan d' chantillonnage D_s . Pour r gler ce probl me, nous proposons d'estimer plut t l'E M non conditionnelle, $E_{QM}(Y) = E[(\hat{Y} - Y)^2] = E_{D_s}\{E_p[(\hat{Y} - Y)^2 | D_s]\}$, o  $E_{D_s} = E_p E_s$ d finit l'espace pour la distribution  chantillonnale (compte tenu de l' chantillon s lectionn ), et pour toutes les s lections possibles d' chantillons. On notera que $E_p[(\hat{Y} - Y)^2 | D_s]$ peut  tre consid r e comme une variable al atoire $u(D_s)$, de sorte que $E_{QM}(\hat{Y}) = E_{D_s}[u(D_s)]$ d finit son  moyenne pr dicteur   par rapport   la fonction de perte des moindres carr s avec la distribution f_{D_s} pour laquelle l'esp rance E_{D_s} est calcul e. En changeant l'ordre des esp rances, on peut reformuler l'E M non conditionnelle comme suit,

$$E_{QM}(Y) = E_s E_p E_{D_s}[(\hat{Y} - Y)^2 | Y] \\ = E_p E_{D_s} E_s[(\hat{Y} - Y)^2 | Y] \quad (6.1)$$

O  $Y = \{y_i; i \in U\}$. On peut ensuite estimer l'E M non conditionnelle pour l'un ou l'autre des pr dicteurs \hat{Y} , en  stimant son E M de randomisation (voir Pfeiffermann (1993) pour une discussion plus d taill e). L'estimation de l'E M de randomisation pour les divers pr dicteurs pr sente l'avantage bas e sur le plan d' chantillonnage, dans une approche bas e sur le plan d' chantillonnage.

L'estimation des variances de randomisation, pour les estimateurs bas s sur le plan d' chantillonnage, a  t   tudi e abondamment dans la litt rature, et des m thodes

nombreuses et diverses sont couramment utilis es. Toutefois, compte tenu de la structure compliqu e de certains des pr dicteurs  tudi s dans cet article et afin de ne pas nous restreindre   un sch ma d' chantillonnage particulier, nous proposons ci-dessous d'utiliser une proc dure en deux  tapes qui combine un processus d' chantillonnage inverse ( tape 1) et ce que nous pouvons consid rer comme  tant un algorithme de r  chantillonnage selon la m thode du bootstrap ( tape 2). Cette proc dure pr sente un avantage notable : elle est g n rale et s'applique   galement   tous les pr dicteurs. En outre,   la diff rence des autres m thodes d'estimation de la variance couramment utilis es, elle ne requiert pas de connaissance des probabilit s de s lection conjointe par couple $\pi_{ij} = \Pr(i, j \in s)$. Comme nous le mentionnons plus loin, une application valide de la premi re  tape requiert des  chantillons suffisamment gros. Les deux  tapes de la proc dure que nous proposons sont les suivantes :

 tape 1 – G n rer une   pseudo-population   unique en s lectionnant N unit s avec remplacement   partir de l' chantillon original, et avec des probabilit s proportionnelles   w_i , o  N est la taille de la population. Cette  tape est justifi e ci-dessous (voir  galement la remarque 10). Notons Y^{pp} la somme des valeurs y dans cette pseudo-population.

 tape 2 – S lectionner ind pendamment un grand nombre B d' chantillons bootstrap   partir de la pseudo-population g n r e   l' tape 1, avec le m me sch ma d' chantillonnage que pour la s lection de l' chantillon original, puis en  stimant \hat{Y} qui repr sente l'un ou l'autre des pr dicteurs, et d notons le pr dicteur obtenu pour l' chantillon bootstrap b par \hat{Y}_b^{pp} . Ensuite, calculons,

$$E_p E_{D_s}[(\hat{Y} - Y)^2] = \frac{1}{B} \sum_{b=1}^B (Y^{pp} - \hat{Y}_b^{pp})^2. \quad (6.2)$$

L'efficacit  de l'estimateur (6.2) pour estimer l'E M de randomisation d pend  videmment du   degr  de justesse   de la pseudo-population g n r e   l' tape 1 par rapport   la population r elle d'o  est tir  l' chantillon original. On peut v rifier le degr  de justesse des deux populations en partie en notant que la distribution marginale de $y_i | x_i$ dans la pseudo-population est la m me que dans la population originale. Pour le constater, il suffit de remarquer que la probabilit  de s lection C_{w_i} pour chaque   tirage  , o  $C = 1 / \sum_{i=1}^N w_i$. Notons $f^{(Y^{pp} | x_i)}$ comme  tant la distribution marginale de la pseudo-population, et nous trouvons   l'aide de (2.2) et de (2.5a),

Remarque 6. Si le modèle ajusté aux données de l'échantillon est une régression linéaire avec une valeur à l'origine et une variance résiduelle constante, le prédicteur $\hat{Y}_{2, \text{Reg}}$ défini par (4.8) et le prédicteur \hat{Y}_3 diffèrent en ce que $\hat{Y}_{2, \text{Reg}}$ utilise un estimateur convergent pour les coefficients de régression qui définissent l'approximation linéaire du modèle, et valable pour le complément de l'échantillon, tandis qu'avec \hat{Y}_3 , les coefficients de régression sont estimés par les moindres carrés ordinaires, ce qui revient à estimer l'approximation linéaire du modèle de l'échantillon.

Enfin, plutôt que de prévoir seulement les valeurs du complément de l'échantillon, comme avec les prédicteurs précédents, on pourrait plutôt prévoir toutes les valeurs de la population d'après leurs espérances estimées, en vertu du régression linéaire avec une valeur à l'origine et une variance résiduelle constante, l'application de (2.5b) donne,

$$\begin{aligned} \beta &= \arg \min_{\beta} E^p(Y^k - \mathbf{x}_k^T \beta)^2 \\ &= \arg \min_{\beta} \frac{E^s(w_k) \beta}{E^s[w_k(Y^k - \mathbf{x}_k^T \beta)^2]} \end{aligned} \quad (4.13)$$

En estimant l'espérance de l'échantillon au numérateur de (4.13) par la méthode de l'échantillon correspondant (application de la méthode des moments) et en minimisant la moyenne de l'échantillon par rapport à β , on obtient l'estimateur pondéré de probabilité familial $\hat{B}^{pw} = (X^{[s]T} W^s X^{[s]})^{-1} (X^{[s]T} W^s Y^s)$, où $(X^{[s]}, Y^s) = ([x_1 \dots x_n]^T, (y_1 \dots y_n)^T)$ et $W^s = \text{Diag}[w_1 \dots w_n]$. Soit $\mathbf{x}_k^T = (1, \tilde{\mathbf{x}}_k^T)$. Si on calcule $\hat{E}^p(Y^k) = \mathbf{x}_k^T \hat{B}^{pw} = B_0 + \tilde{\mathbf{x}}_k^T \hat{B}^{pw}$ et on somme sur toutes les valeurs de la population, on obtient l'estimateur familial de régression généralisée (GREG) (Sarndal 1980),

$$\hat{Y}_{\text{GREG}} = N \frac{\sum_{i \in s} w_i Y_i}{\sum_{i \in s} w_i} + \tilde{B}^{pw} \left[\tilde{X}(p) - N \frac{\sum_{i \in s} w_i \mathbf{x}_i}{\sum_{i \in s} w_i} \right] \quad (4.14)$$

Remarque 7. En considérant l'estimation de Y comme un problème prévisionnel, l'utilisation du prédicteur $\hat{Y}_{2, \text{Reg}}$ dans (4.8) requiert la prévision de $(N - n)$ valeurs, tandis que l'utilisation d'un estimateur GREG requiert la prévision de N valeurs. Ainsi, dans une situation où le modèle de l'échantillon et le modèle de population peuvent être relativement bien approximatés par des modèles de régression linéaire avec des valeurs à l'origine (mais

pour les deux modèles), on pourrait s'attendre à ce que pour

des fractions d'échantillonnage relativement importantes n/N , le prédicteur $\hat{Y}_{2, \text{Reg}}$ soit supérieur (pour des résultats empiriques, voir la section 7).

5. EXEMPLES

5.1 Prédiction sans variable concomitante

Soit $\mathbf{x}_i = 1$ pour tous les i . D'après (3.2),

$$\begin{aligned} \hat{Y} &= \sum_{i \in s} Y_i + \sum_{j \in s} \hat{E}^c(Y_j) - \sum_{i \in s} Y_i \\ &\quad + (N - n) \hat{E}^s \left(\frac{\hat{E}^s(w_j) - 1}{w_j - 1} Y_j \right). \end{aligned} \quad (5.1)$$

En estimant les deux espérances de l'échantillon du côté droit de (5.1) d'après leur moyenne d'échantillon respective, on obtient l'estimateur,

$$\begin{aligned} \hat{Y}_{\text{El}} &= \sum_{i \in s} Y_i + (N - n) \frac{1}{l} \sum_{i \in s} \frac{w_i - 1}{w_i} Y_i \\ &= \sum_{i \in s} Y_i + \frac{\sum_{i \in s} (w_i - 1)}{(N - n)} \sum_{i \in s} (w_i - 1) Y_i. \end{aligned} \quad (5.2)$$

Dans (5.2), $\sum_{i \in s} (w_i - 1) Y_i$ est un « estimateur de Horvitz-Thompson » de $\sum_{j \in s} Y_j$. Le multiplicateur $(N - n) / \sum_{i \in s} (w_i - 1)$ est une « correction de type Hajek » visant à contrôler la variabilité des poids de l'échantillon. On notera que \hat{Y}_{El} est un cas spécial du prédicteur $\hat{Y}_{2, \text{Reg}}$ défini à (4.8), et obtenu en choisissant $\mathbf{x}_i = 1$ pour toutes les valeurs i . C'est également un cas spécial du prédicteur \hat{Y}_3 si on estime $\hat{E}^s(Y_j) = \bar{y} = \sum_{i \in s} Y_i / n$. Pour les plans d'échantillonnage tels que $\sum_{i \in s} w_i = N$ pour tous les s , ou si on estime $\hat{E}^s(w_i) = N / n$, le prédicteur \hat{Y}_{El} se réduit à l'estimateur familial d'Horvitz-Thompson pour le total de population, soit $\hat{Y}_{\text{H-T}} = \sum_{i \in s} w_i Y_i$.

Tout comme avec l'estimateur GREG mentionné à la section 4, au lieu de prévoir le total du complément de l'échantillon $Y_c = \sum_{j \in s} Y_j$ et en utilisant, le prédicteur \hat{Y}_{El} , on pourrait prévoir toutes les valeurs y dans la population, estimant leur espérance d'après le modèle de population. Selon (2.5b), $E^p(Y_i) = E^s(w_i Y_i) / E^s(w_i)$. En estimant les deux espérances d'échantillon d'après les moyennes de l'échantillon correspondantes, on obtient l'estimateur de Hajek familial,

$$\hat{Y}_{\text{Hajek}} = \sum_{k=1}^N \hat{E}^p(Y^k) = N \hat{E}^s \left(\frac{\hat{E}^s(w_i)}{w_i Y_i} \right) \sum_{i \in s} \frac{w_i}{Y_i} \quad (5.3)$$

Le prédicteur X_2 présente un cas spécial intéressant quand le modèle de travail posé pour le complément de l'échantillon est linéaire, avec une valeur à l'origine et une variance constante. Soit $x_i' = (1, \bar{x}_i')$. Comme on peut le vérifier aisément, l'estimateur prend, dans ce cas-ci, la forme,

$$\hat{Y}_2, \text{Reg} = \sum_{j \in S} y_j + \hat{Y}_c + \hat{B}_c [\bar{X}(c) - \bar{X}_c] \quad (4.8)$$

où $\bar{X}(c) = \sum_{j \in S} \bar{x}_j$, $(Y_c, \bar{X}_c) = [(N-n)/\sum_{j \in S} (w_j - 1)] [\sum_{j \in S} (w_j - 1)(y_j, \bar{x}_j)]$ et B_c est l'estimateur pondéré de probabilité du coefficient vectoriel de \bar{x}_i , mais avec les poids $(w_j - 1)$, au lieu de w_j .

Remarque 5. Le prédicteur \hat{Y}_2, Reg peut être obtenu comme cas spécial des prédicteurs esthétiques proposés par Brewer (1999). Toutefois, nous devons souligner que le développement des prédicteurs esthétiques et le calcul de leurs EQM présupposent explicitement un échantillonnage non informatif.

Une propriété importante de \hat{Y}_2, Reg est que, dans des conditions générales, il assure la convergence par rapport au plan d'échantillonnage pour X , peu importe le modèle du complément d'échantillon véritable (voir le lemme 1 ci-dessous). Pour de nombreux analystes, la « convergence par rapport au plan d'échantillonnage » est une exigence essentielle pour tout prédicteur; voir la discussion dans Hansen, Madow et Tepping (1983) et Samdal (1980). Le lemme 1, qui suit, définit des conditions dans lesquelles le prédicteur plus général \hat{X}_2 de (4.7) assure la convergence par rapport au plan d'échantillonnage de X .

Lemme 1. Le prédicteur \hat{X}_2 assure la convergence par rapport au plan d'échantillonnage pour X si le modèle de travail utilisé pour le calcul de \hat{Y}_2 répond aux conditions suivantes : i- $C_\beta(x)$ a une valeur à l'origine; ii- $C_\beta(x)$ est différentiable par rapport à β dans le voisinage de β_2 et

iii- $v(x) = \text{constante}$.

Preuve : D'après (4.6) et la condition iii, $\beta_2 = \arg \min_{\beta} [\sum_{j \in S} (w_j - 1) y_j - C_\beta(x)]$ et d'après la condition i, $C_\beta(x) = \beta_0 + C_{\beta, \beta_2}(x)$, de sorte qu'en vertu de la condition ii, $\partial/\partial \beta_0 \{ \sum_{j \in S} (w_j - 1) [y_j - C_\beta(x)] \}_{\beta=\beta_2} = 0$, ce qui donne $\sum_{j \in S} (w_j - 1) [y_j - C_{\beta_2}(x)] = 0$, ou,

$$\sum_{j \in S} w_j y_j = \sum_{j \in S} y_j + \sum_{j \in S} w_j C_{\beta_2}(x_j) - \sum_{j \in S} C_{\beta_2}(x_j). \quad (4.9)$$

La preuve se termine en notant que dans des conditions de régularité moyennes, $\sum_{j \in S} w_j y_j$ assure la convergence par rapport au plan d'échantillonnage pour X , et $\sum_{j \in S} w_j C_{\beta_2}(x_j)$ est la convergence par rapport au plan d'échantillonnage pour $\sum_{j=1}^N C_{\beta_2}(x_j)$. Ainsi, le membre droit de l'équation (4.9) converge en probabilité vers X_2 , tandis que le membre gauche converge en probabilité vers X .

Il est important de souligner de nouveau que le lemme 1 ne repose pas sur l'hypothèse que le modèle de travail est le bon modèle de complément de l'échantillon.

Pour utiliser les prédicteurs \hat{X}_1 et \hat{X}_2 , il faut spécifier le modèle du complément de l'échantillon. L'étape suivante consiste donc à développer un autre prédicteur qui requiert uniquement l'identification et l'estimation du modèle d'échantillonnage. L'approche menant à ce prédicteur est l'analogie, pour le complément de l'échantillon, de la « méthode avec correction du biais », proposée par Chambers et coll. (2003). Le prédicteur proposé est basé sur la relation suivante :

$$\sum_{j \in S} E_c(y_j | x_j) = \sum_{j \in S} E_s(y_j | x_j) + (N-n) \left[\frac{1}{N-n} \sum_{j \in S} E_c \{ [y_j - E_s(y_j | x_j)] | x_j \} \right] \quad (4.10)$$

$$+ (N-n) \left[\frac{1}{N-n} \sum_{j \in S} E_s \{ y_j - E_s(y_j | x_j) \} \right]$$

où dans la deuxième rangée nous avons remplacé la moyenne (pour le complément de l'échantillon) des espérances conditionnelles $E_s(y_j | x_j)$ par son espérance pour la distribution des valeurs de X sur le complément de l'échantillon (n indique la taille de l'échantillon). D'après

(2.9),

$$E_c[y_j - E_s(y_j | x_j)] = E_s \left\{ \left[\frac{E_s(w_j) - 1}{w_j - 1} y_j - E_s(y_j | x_j) \right] \right\} \quad (4.11)$$

signifie que la moyenne du complément de l'échantillon à la deuxième rangée de (4.10) peut être estimée par $\hat{M}_c = 1/n \sum_{j \in S} \{ (w_j - 1) [y_j - E_s(y_j | x_j)] \}$, où $\hat{w}_s = \sum_{j \in S} w_j / n$. Le prédicteur proposé prend donc la forme

$$\hat{Y}_2 = \sum_{j \in S} y_j + \sum_{j \in S} \hat{E}_s(y_j | x_j) + (N-n) \hat{M}_c \quad (4.12)$$

où $\hat{E}_s(y_j | x_j)$ est estimé à partir des données de l'échantillon. L'utilisation de \hat{Y}_2 requiert seulement l'identification et l'estimation de la régression de l'échantillon $E_s(y_j | x_j)$, réalisables à l'aide des techniques classiques de régression. Qui plus est, dans des conditions peu sévères, \hat{Y}_2 est la convergence par rapport au plan d'échantillonnage pour X , même si l'espérance $E_s(y_j | x_j)$ est mal spécifiée. Cette propriété découle du fait que $\sum_{j \in S} \hat{E}_s(y_j | x_j)$ est la convergence par rapport au plan d'échantillonnage pour $\sum_{j \in S} E_s(y_j | x_j)$, et que $(N-n) \hat{M}_c$ est également convergent pour $M_c = \sum_{j \in S} [y_j - E_s(y_j | x_j)]$.

4. PRÉVISION SEMI PARAMÉTRIQUE DES TOTAUX DE POPULATION FINIE

Supposons que le modèle du complément d'échantillon prend la forme suivante :

$$\begin{aligned} y_j &= C_\beta(\mathbf{x}_j) + \varepsilon_j, \\ E_c(\varepsilon_j | \mathbf{x}_j) &= 0, E_c(\varepsilon_j^2 | \mathbf{x}_j) = \sigma^2 v(\mathbf{x}_j), \\ E_c(\varepsilon_j \varepsilon_k | \mathbf{x}_j, \mathbf{x}_k) &= 0, k \neq j \end{aligned} \quad (4.1)$$

où $C_\beta(\mathbf{x})$ est une fonction connue (probablement non linéaire) de \mathbf{x} , qui dépend d'un paramètre vectoriel inconnu β . On suppose que les variances $\sigma^2 v(\mathbf{x}_j)$ sont connues, sauf pour σ^2 .

Remarque 3. Dans des applications réelles, on peut identifier le modèle (4.1) selon une procédure en deux

étapes, en utilisant l'égalité $E_c(y_j | \mathbf{x}_j) = E_s(r_j y_j | \mathbf{x}_j)$, avec $r_j = (w_j - 1) / E_s[(w_j - 1) | \mathbf{x}_j]$ (qui découle de l'équation 2.9). Estimons d'abord $E_s(w_j | \mathbf{x}_j)$, et ensuite r_j , par régression de w_j par rapport à \mathbf{x}_j , en utilisant les données de l'échantillon. Soit $\hat{r}_j = (w_j - 1) / \hat{E}_s(w_j | \mathbf{x}_j) - 1$ et la transformée $y_j^* = \hat{r}_j y_j$. En deuxième lieu, étudions la relation dans l'échantillon entre y_j^* et \mathbf{x}_j , afin d'identifier la forme de $C_\beta(\mathbf{x}_j)$. Voir Pfeffermann et Sverchov (1999, 2003a), qui présentent des exemples d'estimation de $E_s(w_j | \mathbf{x}_j)$. On peut appliquer une procédure similaire afin d'identifier la fonction de variance $v(\mathbf{x}_j)$ à l'aide des valeurs résiduelles empiriques $\hat{\varepsilon}_j = y_j^* - \hat{E}_s(\hat{r}_j y_j^* | \mathbf{x}_j)$.

La fonction $C_\beta(\mathbf{x}_j)$ dans (4.1) avec le paramètre vectoriel vrai β satisfait, pour toutes les valeurs $j \notin s$,

$$\begin{aligned} C_\beta(\mathbf{x}_j) &= \arg \min_{E_c} \left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \middle| \mathbf{x}_j \right) \\ &= \arg \min_{E_s} \left(r_j \frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right). \end{aligned} \quad (4.2)$$

(La deuxième égalité découle de (2.9). Par conséquent, en substituant l'espérance de l'échantillon à l'extérieur des grandes parenthèses par la moyenne de l'échantillon (une application directe de la méthode des moments) et en estimant r_j par \hat{r}_j (voir la remarque 3), on peut estimer le vecteur β comme suit :

$$\hat{\beta}_1 = \arg \min_{\beta} \sum_{j \in s} \left(\hat{r}_j \frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right). \quad (4.3)$$

Le prédicteur du total de la population prend alors la forme,

$$(4.4)$$

$$\hat{Y}_1 = \sum y_i + \sum_{j \notin s} \hat{C}_{\beta_1}(\mathbf{x}_j).$$

Par ailleurs, il découle de (4.1) que,

$$E_c \left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \middle| \mathbf{x}_j \right) = E_c \left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right)$$

$$= E_s \left(\frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \right)$$

$$= E_s \left(\frac{[E_s(w_j | \mathbf{x}_j) - 1] \left[\frac{E_s(w_j | \mathbf{x}_j) - 1}{w_j - 1} - 1 \right] \left[\frac{E_s(w_j | \mathbf{x}_j) - 1}{w_j - 1} - 1 \right] \right)$$

où l'espérance à la droite est établie par rapport à la distribution conjointe de (y_j, \mathbf{x}_j) . Par conséquent, on peut estimer β comme suit,

$$\hat{\beta}_2 = \arg \min_{\beta} \sum_{j \in s} (w_j - 1) \frac{[y_j - C_\beta(\mathbf{x}_j)]^2}{v(\mathbf{x}_j)} \quad (4.6)$$

car $\hat{E}_s(w_j) = \text{constante}$. Le prédicteur de Y avec β estimé par $\hat{\beta}_2$ est donc,

$$\hat{Y}_2 = \sum y_i + \sum_{j \notin s} C_{\hat{\beta}_2}(\mathbf{x}_j). \quad (4.7)$$

Remarque 4. Un avantage important du prédicteur \hat{Y}_2 par rapport au prédicteur \hat{Y}_1 est qu'il ne requiert pas l'identification et l'estimation de l'espérance $w(\mathbf{x}) = E_s(\mathbf{x})$. Par ailleurs, dans les situations où cette espérance peut être estimée adéquatement, il est probable que le prédicteur \hat{Y}_1 est plus exact, car les poids $r_i = (w_i - 1) / [E_s(w_i | \mathbf{x}_i) - 1]$ seront souvent moins variables que les poids $(w_i - 1)$. Cela est dû au fait que les poids r_j tiennent compte seulement de l'effet net du processus d'échantillonnage sur la distribution conditionnelle cible $f_c(y_j | \mathbf{x}_j)$, alors que les poids $(w_j - 1)$ tiennent compte de l'effet du processus d'échantillonnage sur la distribution conjointe $f_c(y_j, \mathbf{x}_j)$. En particulier, quand w_i est une fonction déterministe de \mathbf{x}_i , tel que $w_i = w(\mathbf{x}_i)$, le processus d'échantillonnage est non informatif et $f_c(y_j | \mathbf{x}_j) = f_s(y_j | \mathbf{x}_j) = \tilde{f}_c(y_j | \mathbf{x}_j)$. Dans ce cas-ci, l'estimateur $\hat{\beta}_1$ (mais non $\hat{\beta}_2$) coïncide avec l'estimateur GLS (moindres carrés généralisés) optimal d'échantillon. (Pour les données analysées à la section 7, la variance empirique des poids r_i est 1.36, tandis que la variance empirique des poids w_i est 2.66.) À l'opposé, quand les poids d'échantillonnage w_i sont indépendants de \mathbf{x}_i , les estimations de $\hat{\beta}_1$ et $\hat{\beta}_2$, et donc des prédicteurs \hat{Y}_1 et \hat{Y}_2 , sont égales car $w(\mathbf{x}_i) = \text{constante}$.

Il découle de (2.4) que

$$a) \quad E^s(\mathbf{u}_i | \mathbf{v}_i) = \frac{E^d(\pi_i | \mathbf{v}_i)}{1};$$

$$b) \quad E^d(\mathbf{u}_i) = \frac{E^s(w_i)}{E^s(\mathbf{u}_i)};$$

$$c) \quad E^s(w_i) = \frac{E^d(\pi_i)}{1}.$$

Pour une discussion détaillée de la distribution échantillonnale avec des illustrations, voir Pfeffermann et coll. (1998).

2.2 Répartition du complément de l'échantillon
Comme dans (2.1), nous définissons comme suit la *fdp* conditionnelles pour les unités à l'extérieur de l'échantillon :

$$f^c(y_i | \mathbf{x}_i) = f^d(y_i | \mathbf{x}_i, I_i = 0)$$

$$= \frac{\Pr(I_i = 0 | y_i, \mathbf{x}_i) f^d(y_i | \mathbf{x}_i)}{\Pr(I_i = 0 | \mathbf{x}_i)}.$$

Les relations (2.2)-(2.5) et l'égalité $\Pr(I_i = 0 | \mathbf{u}_i, \mathbf{v}_i) = 1 - \Pr(I_i = 1 | \mathbf{u}_i, \mathbf{v}_i) = 1 - E^d(\pi_i | \mathbf{u}_i, \mathbf{v}_i)$ donnent lieu aux représentations suivantes de la répartition du complément de l'échantillon pour des paires générales de variables aléatoires vectorielles $(\mathbf{u}_i, \mathbf{v}_i)$:

$$f^c(\mathbf{u}_i | \mathbf{v}_i) = \frac{E^d[(1 - \pi_i) | \mathbf{u}_i, \mathbf{v}_i] f^d(\pi_i | \mathbf{u}_i, \mathbf{v}_i)}{E^d[(1 - \pi_i) | \mathbf{v}_i]}$$

$$= \frac{E^d[(1 - \pi_i) | \mathbf{u}_i, \mathbf{v}_i] E^d[\pi_i | \mathbf{u}_i, \mathbf{v}_i]}{E^d[\pi_i | \mathbf{v}_i]} f^s(\mathbf{u}_i | \mathbf{v}_i) \quad (2.7)$$

$$f^c(\mathbf{u}_i | \mathbf{v}_i) = \frac{E^s[(w_i - 1) | \mathbf{u}_i, \mathbf{v}_i] f^s(\mathbf{u}_i | \mathbf{v}_i)}{E^s[(w_i - 1) | \mathbf{v}_i]}.$$

(L'équation (2.8) découle de l'application de (2.5a) avec la deuxième expression de (2.7)). En outre, selon (2.8) et la première équation de (2.7),

$$E^c(\mathbf{u}_i | \mathbf{v}_i) = \frac{E^d[(1 - \pi_i) | \mathbf{u}_i, \mathbf{v}_i]}{E^s[(w_i - 1) | \mathbf{v}_i]} = \frac{E^d[(1 - \pi_i) | \mathbf{v}_i]}{E^s[(w_i - 1) | \mathbf{v}_i]}.$$

Remarque 2. Dans les applications pratiques, la fraction d'échantillonnage est souvent très faible et par conséquent les probabilités de sélection de l'échantillon sont faibles pour au moins la plupart des unités de population. Si $\pi_i > \delta$ avec une probabilité de 1,

$$f^c(\mathbf{u}_i | \mathbf{v}_i) = \frac{E^d[(1 - \pi_i) | \mathbf{u}_i, \mathbf{v}_i] f^d(\pi_i | \mathbf{u}_i, \mathbf{v}_i)}{E^d[(1 - \pi_i) | \mathbf{v}_i]}$$

$$= f^d(\mathbf{u}_i | \mathbf{v}_i) +$$

$$\frac{E^d[(\pi_i | \mathbf{u}_i, \mathbf{v}_i) - \pi_i | \mathbf{u}_i, \mathbf{v}_i] f^d(\pi_i | \mathbf{u}_i, \mathbf{v}_i)}{E^d[(1 - \pi_i) | \mathbf{v}_i]}$$

$$= f^d(\mathbf{u}_i | \mathbf{v}_i) (1 + \Delta) \quad (2.10)$$

où $-\delta < \Delta < \delta/(1 - \delta)$. Il découle de (2.10) que pour les valeurs de δ suffisamment petites, la différence entre la *fdp* de la population et la *fdp* du complément de l'échantillon est très faible, ce qui n'est guère surprenant.

3. PRÉVISION OPTIMALE DES TOTAUX DE POPULATION FINIE

Soit $Y = \sum_{i=1}^N y_i$ qui définit le total de population. Le problème est le suivant : comment prévoir Y d'après les données de l'échantillon et peut-être les valeurs, dans la population, des variables auxiliaires? Dénons l'« information sur le plan » disponible pour la prévision par $D_s = \{(y_i, w_i), i \in s; (\mathbf{x}_j, I_j), j = 1 \dots N\}$, et $Y = Y(D_s)$ définit le prédicteur. L'EQM de Y par rapport à la *fdp* de la population, compte tenu de D_s , est :

$$\begin{aligned} \text{EQM}(Y | D_s) &= E^d[Y - Y | D_s] \\ &= E^d\{Y - E^d(Y | D_s) | D_s\} \\ &= E^d\{Y - E^d(Y | D_s) + V^d(Y | D_s)\} \end{aligned}$$

car $[Y - E^d(Y | D_s)]$ est fixé pour une valeur D_s donnée. Il découle de (3.1) que $\text{EQM}(Y | D_s)$ est minimisé quand $Y = E^d(Y | D_s)$. Cette dernière espérance peut être déterminée comme suit :

$$\begin{aligned} E^d(Y | D_s) &= \sum_{i=1}^N E^d(y_i | D_s, I_i = 1) + \sum_{j \notin s} E^d(y_j | D_s, I_j = 0) \\ &= \sum_{j \notin s} y_j + \sum_{j \in s} E^d(y_j | \mathbf{x}_j) \\ &= \sum_{j \in s} y_j + \sum_{j \notin s} E^c(y_j | \mathbf{x}_j) \end{aligned} \quad (3.2)$$

où, dans la dernière égalité, nous supposons que les valeurs y_j pour $j \notin s$ et D_s ne sont pas corrélées, compte tenu de \mathbf{x}_j . Le problème prévisionnel se réduit donc à estimer les espérances $E^c(y_j | \mathbf{x}_j)$. À la section 4, nous étudions l'estimation semi paramétrique de ces espérances.

2. DISTRIBUTIONS DE L'ÉCHANTILLON ET DU COMPLÈMENT DE L'ÉCHANTILLON

2.1 Distribution échantillonnaie

Supposons que les valeurs de la population $\{Y, X\} = \{Y_1, \dots, Y_N, X_1, \dots, X_N\}$ sont des réalisations aléatoires, selon la fonction de densité de probabilité (*fdp*) conditionnelle $f_p(y_i | x_i)$, qui peut être discrète ou continue. On suppose que les valeurs y sont scalaires, mais les valeurs x peuvent être des vecteurs. Nous considérons l'échantillonnage à un seul degré, avec une probabilité d'inclusion dans l'échantillon de $\pi_i = \Pr(I_i = s) = g(Y, X, Z, i)$ pour une fonction g donnée, où Z définit les valeurs de la population des variables utilisées pour le processus d'échantillonnage. On notera que les valeurs y sont aléatoires, et nous pouvons également considérer les variables du plan comme étant aléatoires, de sorte que les valeurs g sont également aléatoires. Soit $I_i = 1$ si $i \in s$ et $I_i = 0$, si $i \notin s$. La *fdp d'échantillon* marginale conditionnelle est définie comme suit :

$$\begin{aligned} f_s(y_i | x_i) &= f(y_i | x_i, I_i = 1) \\ &= \frac{\Pr(I_i = 1 | y_i, x_i) f_p(y_i | x_i)}{\Pr(I_i = 1 | x_i)} \end{aligned} \quad (2.1)$$

la deuxième égalité étant obtenue par application du théorème de Bayes. On notera que $\Pr(I_i = 1 | y_i, x_i)$ n'est pas nécessairement identique à la probabilité de sélection d'échantillon réelle $\pi_i = g(Y, X, Z, i)$ (voir la remarque 1 ci-dessous). Par conséquent, il découle de (2.1) que les *fdp* de la population d'échantillon sont différentes, à moins que $\Pr(I_i = 1 | y_i, x_i) = \Pr(I_i = 1 | x_i)$ pour toutes les valeurs y_i . Quand la distribution échantillonnaie diffère de la répartition de la population, elle acquiert un caractère informatif, et le schéma d'échantillonnage ne peut être ignoré dans le processus d'inférence.

Remarque 1. Il est important de souligner que la définition et l'utilisation de la répartition de sélection ne présupposent pas que les probabilités de sélection d'échantillon sont fonction uniquement des couples (y_i, x_i) . Comme nous l'avons mentionné plus tôt et comme nous l'avons indiqué en exprimant les probabilités de sélection sous la forme $\pi_i = g(Y, X, Z, i)$, les probabilités de sélection réelles peuvent dépendre de toutes les valeurs de la population (Y, X, Z) . Toutefois, comme l'ont démontré Pfeffermann et Sverchkov (1999), $E_p(\pi_i | y_i, x_i) = \Pr(I_i = 1 | y_i, x_i)$. Ainsi, même si les probabilités de sélection peuvent dépendre de toutes les valeurs de population (Y, X, Z) , pour des valeurs (y_i, x_i) données, elles sont égales à $\Pr(I_i = 1 | y_i, x_i)$ « en moyenne ». En fait, π_i peut ne pas dépendre du tout de

$$f_s(\mathbf{u} | \mathbf{v}_i) = \frac{E_p(\pi_i | \mathbf{v}_i)}{E_p(\pi_i | \mathbf{v}_i) f_p(\mathbf{u} | \mathbf{v}_i)} \quad (2.2)$$

$$f_p(\mathbf{u} | \mathbf{v}_i) = \frac{E_s(w_i | \mathbf{v}_i)}{E_s(w_i | \mathbf{v}_i) f_s(\mathbf{u} | \mathbf{v}_i)} \quad (2.3)$$

$$E_p(\mathbf{u} | \mathbf{v}_i) = \frac{E_s(w_i | \mathbf{v}_i)}{E_s(w_i | \mathbf{v}_i)} \quad (2.4)$$

Dans les pages suivantes, nous considérons les probabilités π_i comme des réalisations aléatoires de la variable aléatoire $g(Y, X, Z, i)$. Soit $w_i = 1/\pi_i$, qui définit le poids d'échantillonnage de l'unité i . Les relations suivantes, établies par Pfeffermann et Sverchkov (1999), sont valables pour des couples généraux de variables aléatoires vectorielles $(\mathbf{u}_i, \mathbf{v}_i)$, où E_p et E_s définissent les espérances en vertu des *fdp* de population et d'échantillon, respectivement (un cas spécial étant $\mathbf{u}_i = y_i$, $\mathbf{v}_i = x_i$).

De la discussion précédente, on ne doit pas conclure que π_i n'est jamais une fonction de (y_i, x_i) seulement. Un exemple classique, dans ce dernier cas, est l'échantillonnage rétrospectif. Ainsi, dans une étude cas-témoin, les probabilités de sélection des cas et des témoins dépendent habituellement uniquement des valeurs y . Dans l'étude empirique que nous présentons ici, nous utilisons un ensemble de données réelles, dans lequel l'échantillon est tiré d'un échantillon stratifié disproportionné, les frontières des strates étant définies par les valeurs de la variable dépendante.

toutes les valeurs y_i , et être uniquement une fonction toujours égale à $\Pr(I_i = 1 | y_i, x_i)$. La raison pour laquelle l'espérance peut dépendre de y_i dans ce cas-ci est que la valeur Z peut être corrélée avec y . Par exemple, l'Enquête canadienne sur le milieu de travail et les employés de 1999 utilise un échantillon stratifié disproportionné, les strates étant définies par région, activité et taille des effectifs. L'information sur la taille des effectifs est tirée des données fiscales de 1998 et 1999, Hiddington et Lavallée (2000) pour plus de détails. Quand on modélise les listes de paie de 1999 par rapport aux nombre d'employés, on constate que le plan d'échantillonnage est de type informatif, ce qui s'explique par le fait que la stratification est basée en partie sur la taille obtenue à partir des dossiers fiscaux de l'année précédente, lesquels sont corrélés avec les listes de paie de l'année subséquente. Voir Fuller (2003) pour une analyse détaillée.

Prévision des totaux de population finie basée sur la distribution échantillonnale

MICHAÏL SVERCHKOV et DANNY PFEFFERMANN¹

RÉSUMÉ

Dans cet article, nous étudions l'utilisation de la distribution échantillonnale pour prévoir les totaux de population finie à l'aide d'un échantillonnage à un seul degré. Les prédicteurs proposés emploient les valeurs échantillonnées de la variable d'enquête cible, les poids d'échantillonnage et les valeurs (peut-être connues) des variables auxiliaires dans la population. Nous résolvons le problème prévisionnel en estimant l'espérance des valeurs de l'étude pour les unités à l'extérieur de l'échantillon, en fonction de l'espérance correspondante selon la distribution échantillonnale et les poids d'échantillonnage. L'erreur quadratique moyenne de la prévision est estimée par la combinaison d'une procédure d'échantillonnage inverse et d'une méthode de rééchantillonnage. Un résultat intéressant de la présente analyse est que plusieurs estimateurs familiaux, d'usage courant, sont en fait des cas spéciaux de l'approche proposée, et celle-ci leur en donne donc une nouvelle interprétation. L'efficacité des nouveaux prédicteurs et de quelques prédicteurs couramment utilisés est évaluée et comparée par simulation de Monte Carlo avec un ensemble de données réelles.

MOTS CLÉS : Méthode du bootstrap; convergence par rapport au plan d'échantillonnage; échantillonnage informatif; distribution du bootstrap; convergence par rapport au plan d'échantillonnage; échantillonnage informatif.

1. INTRODUCTION

La distribution échantillonnale est la répartition paramétrique des valeurs résultantes des unités incluses dans l'échantillon. La répartition est différente de la répartition de la population si les probabilités de sélection de l'échantillon sont corrélées avec les valeurs de la variable étudiée, même en cas de conditionnement de valeurs des variables concomitantes incluses dans le modèle de la population. Elle diffère également de la répartition (selon le plan d'échantillonnage) de la randomisation qui tient compte de toutes les sélections possibles de l'échantillon, pour des valeurs de population maintenues fixes. La distribution échantillonnale est définie et discutée avec exemples à l'appui dans Pfeffermann, Krieger et Rinott (1998), et elle est étudiée en détail par Pfeffermann et Sverchkov (1999), qui l'utilisent pour estimer des modèles de régression linéaire. Pfeffermann (1997) utilise la répartition de l'échantillonage pour tester les fonctions de répartition de la population, et Pfeffermann et Sverchkov (2003a) discutent de son utilisation pour l'ajustement des modèles linéaires généralisés. Chambers, Dorfman et Sverchkov (2003) utilisent la répartition de l'échantillonnage pour l'estimation non paramétrique des modèles de régression, et Kim (2002) et Pfeffermann et Sverchkov (2003b) l'appliquent à des problèmes d'estimation à l'égard des petites régions.

Dans cet article, nous étudions l'utilisation de la répartition de l'échantillonnage pour prévoir les totaux de population finie, en vertu d'un échantillonnage à un seul

degré. Nous faisons l'hypothèse que les valeurs résultantes pour la population (les valeurs y) sont des réalisations aléatoires d'une distribution donnée qui conditionnent les valeurs connues des variables auxiliaires (les valeurs x). Le problème consiste donc à prévoir la valeur Y totale de la population basée sur les valeurs y de l'échantillon, les poids d'échantillonnage pour les unités dans l'échantillon et les valeurs x de la population. L'utilisation de la distribution échantillonnale permet de conditionner toutes ces valeurs, ce qui n'est pas possible avec une répartition (selon le plan d'échantillonnage) par randomisation, et la prévision de Y est donc équivalente à prévoir les valeurs y pour les unités à l'extérieur de l'échantillon.

Le problème prévisionnel est résolu par l'estimation de l'espérance conditionnelle des valeurs y (compte tenu des valeurs x) pour les unités à l'extérieur de l'échantillon, en fonction de l'espérance de l'échantillon conditionnel (c'est-à-dire l'espérance en vertu de la répartition de l'échantillon) et des poids d'échantillonnage. L'erreur quadratique moyenne de la prévision est estimée par la combinaison d'une procédure d'échantillonnage inverse et d'une méthode de rééchantillonnage. Nous constatons que plusieurs estimateurs familiaux et couramment utilisés, notamment les estimateurs basés sur un plan d'échantillonnage classique, sont en fait des cas spéciaux de la procédure proposée, ce qui leur en donne donc une nouvelle interprétation. L'efficacité des prédicteurs (les nouveaux comme les anciens) est évaluée et comparée par simulation de Monte Carlo avec un ensemble de données réelles.

- ESTEVAO, V.M., et SÄRNDAL, C.-E. (1999). Le recours à des informations auxiliaires pour les estimations par domaines fondées sur le plan de sondage. *Techniques d'enquête*, 25-2, 241-250.
- HARTLEY, H.O. (1959). *Analytic Studies of Survey Data*. Instituto di Statistica, Rome, Volume in honor of Corrado Gini.
- HIDIROGLOU, M.A. (1991). Structure of the Generalized Estimation System (GES). Rapport de Statistique Canada, septembre, 1991.
- HOLT, D., et SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.
- RAO, J.N.K. (1985). Inférence conditionnelle dans les enquêtes par sondage. *Techniques d'enquête*, 11-1, 17-35.
- SÄRNDAL, C.-E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of American Statistical Association*, 84, 405, 266-275.
- SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.
- DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. Dans *New Developments in Survey Sampling*, (Eds. N.L. Johnson et H. Smith). New York: Wiley, Interscience.
- ESTEVAO, V.M., HIDIROGLOU, M.A. et SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 2, 181-204.

BIBLIOGRAPHIE

$$\hat{\mathbf{B}}_{1g} = \left(\sum_{s_g} w_g \mathbf{x}_g \mathbf{x}_g' / c_g \right)^{-1} \sum_{s_g} w_g \mathbf{x}_g y_g / c_g.$$

$$\hat{Y}_{g,HA} = \left(N_g / \hat{N}_g \right) \hat{Y}_{g,HT}^g \hat{Y}_{g,HT} = \sum_{s_g} w_g y_{gk}$$

et

où

Tableau 5
Biais conditionnels des versions par quotient des estimateurs (2.1) à (2.4) et (2.6)

Estimateur	Biais conditionnel
Quotient HT : $\hat{Y}_{d,t\eta}$	$N \tilde{Y}_{U_d}^{U_d} (W_d - W_d) \left(\frac{\tilde{x}_{U_d}}{\tilde{x}_{U_d}} (1 - W_d) \right)$
Quotient stratifié a posteriori : $\hat{Y}_{d,t\eta_2}$	Presque 0
Quotient de rechange HT : $\hat{Y}_{d,t\eta_3}$	$N \tilde{Y}_{U_d}^{U_d} (W_d - W_d) \left(\frac{\tilde{x}_{U_d}}{\tilde{x}_{U_d} - \tilde{x}_{U_d} \tilde{Y}_{U_d}} (W_d - W_d) \right)$
Quotient Hájek : $\hat{Y}_{d,t\eta_4}$	$N \tilde{Y}_{U_d}^{U_d} (W_d - W_d) \left(\frac{\tilde{x}_{U_d}}{W_d \tilde{x}_{U_d} - \tilde{x}_{U_d}} (1 - W_d) \right)$
Quotient de rechange Hájek : $\hat{Y}_{d,t\eta_5}$	Presque 0

Les biais conditionnels présentés à la figure 3 appuient les résultats théoriques présentés au tableau 5. Les trois estimateurs de Hájek sont presque conditionnellement sans biais. L'ordre de grandeur du biais conditionnel de l'estimateur par quotient et de l'estimateur par quotient de rechange HT concorde avec celui du biais conditionnel théorique. Il convient toutefois de souligner que le biais conditionnel de l'estimateur par quotient de rechange HT est plus petit que celui de l'estimateur par quotient HT. En outre, pour les plus grands domaines, ce biais conditionnel est moins prononcé pour l'estimateur par quotient de rechange HT.

Les taux de couverture conditionnels sont présentés aux figures 4a et 4b. Nous constatons que les trois estimateurs de Hájek suivent de près la probabilité de couverture nominale de 95 %. Le taux de couverture de l'estimateur par quotient de rechange HT est raisonnable pour les grands domaines, même si l'est conditionnellement biaisé. Malheureusement, son taux de couverture devient nettement moins bon pour les petits domaines. Le taux de couverture de l'estimateur par quotient HT n'est pas acceptable. Il importe toutefois de souligner que les taux de couverture des estimateurs conditionnellement biaisés s'améliorent à mesure que la taille de l'échantillon réalisée n_d s'approche de la taille d'échantillon de domaine prévue $E(n_d)$.

Sommairement, l'étude en simulation indique que les trois estimateurs de Hájek, c'est-à-dire les estimateurs par quotient stratifié a posteriori de Hájek, par quotient de rechange de Hájek et par quotient de Hájek, sont les meilleurs en ce qui concerne les propriétés conditionnelles et inconditionnelles. Il importe de souligner que, même si l'estimateur par quotient de Hájek est celui qui utilise le moins de données auxiliaires de domaine (il utilise les chiffres de population de domaine N_d), son erreur quadratique moyenne reste raisonnable. L'estimateur par quotient stratifié a posteriori de Hájek est le meilleur en ce qui concerne les propriétés conditionnelles et inconditionnelles.

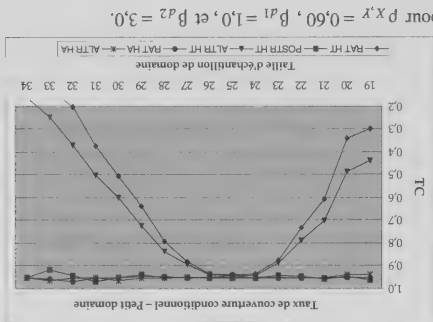
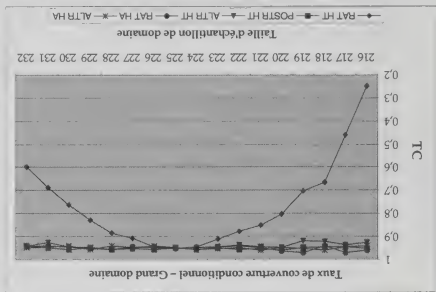
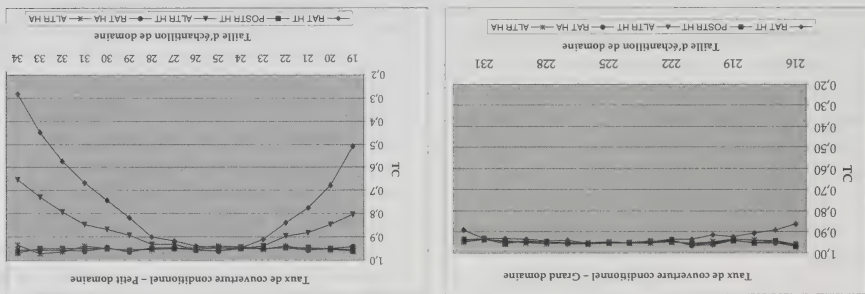
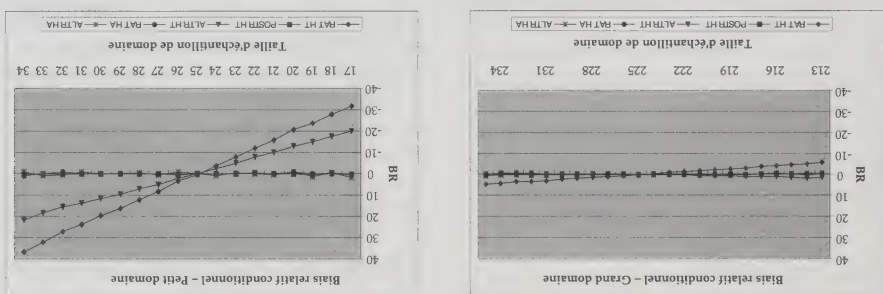
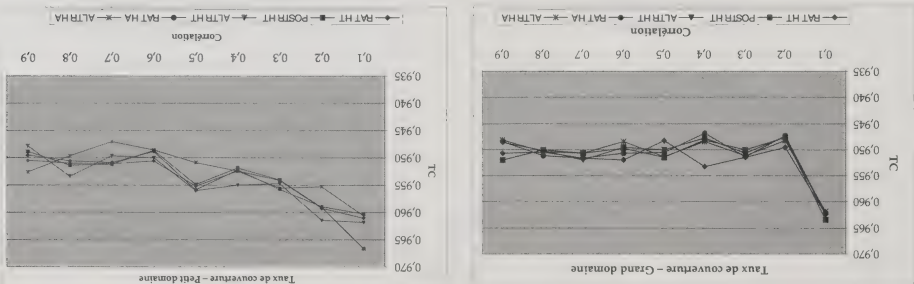
5. CONCLUSION

Nous avons étudié six estimateurs par régression possibles des totaux de domaine, chacun utilisant divers degrés d'information auxiliaire au niveau du domaine et (ou) de la population. Le seul estimateur dont les poids de régression ne dépendent pas du domaine et possèdent la propriété d'additivité est celui d'Horvitz-Thompson $\hat{Y}_{d,t\eta_1}$. Cet estimateur est construit en utilisant l'information auxiliaire au niveau de la population par régression de la variable indépendante dépendant du domaine y_{de} sur le vecteur auxiliaire x_d . Cependant, le biais conditionnel peut être important et les intervalles de confiance connexes peuvent être sous-estimés.

Les estimateurs de type Hájek présentent deux inconvénients : i) ils n'ont pas de propriétés additives et ii) leurs poids de régression dépendent du domaine. Toutefois, ils offrent les meilleures propriétés conditionnelles. Ils sont presque conditionnellement sans biais et les intervalles de confiance conditionnels connexes suivent de près le taux de couverture nominal. Ils ont aussi la plus petite EQM inconditionnelle. L'estimateur de Hájek qui utilise le moins de données auxiliaires au niveau du domaine est $\hat{Y}_{d,t\eta_5}$. Il nécessite les chiffres de population de domaine N_d ($d = 1, \dots, D$) et les totaux de population X . Ses propriétés conditionnelles et inconditionnelles sont raisonnables.

Le meilleur estimateur de Hájek, $\hat{Y}_{d,t\eta_2}$, utilise des données auxiliaires au niveau du domaine. On peut rendre l'estimateur par régression de type Hájek $\hat{Y}_{d,t\eta_1}$ indépendant du domaine en utilisant un ensemble unique de poids de régression comme suit. Supposons que les domaines les plus importants sont $U_g \subseteq U$ ($g = 1, \dots, G$) et que ces domaines sont mutuellement exclusifs et exhaustifs. L'estimateur de Hájek résultant est

$$\hat{Y}_{d,t\eta_2} = \sum_{g=1}^G \left[\hat{Y}_{g,HA} + (X_g - \hat{X}_{g,HA}) \hat{b}_{1g} \right]$$



4.1 Résultats inconditionnels

Nous avons évalué les propriétés inconditionnelles des estimateurs au moyen de deux mesures de performance, à savoir i) la racine carrée de l'erreur quadratique moyenne ou RMSE pour *root mean squared error* et ii) le taux de couverture (TC), définies comme suit :

i. la RMSE est définie comme étant :

$$\sqrt{\frac{\sum_{m=1}^M (\hat{Y}_{(m)}^d - Y_d)^2}{M}} ,$$

où $\hat{Y}_{(m)}^d$ est le total estimé (de type Horvitz-Thompson ou de type Hajék) d'après l'échantillon m et M est le nombre total d'échantillons tirés pour la simulation;

ii. le taux de couverture TC pour un estimateur donné \hat{Y} est défini comme étant le rapport du nombre de fois que l'intervalle de confiance à 95 %

$$\hat{Y}_{(m)}^d \pm 1.96 \sqrt{V(\hat{Y}_{(m)}^d)}$$

contient le total de population réel au nombre de répétitions. Nous utilisons les variances inconditionnelles données par (2.12) et les termes d'erreur du tableau 5 pour estimer les variances requises.

Les quatre graphiques présentés aux figures 1 et 2 résument l'analyse inconditionnelle pour le grand et le petit domaine. Nous montrons aussi l'effet de l'augmentation de la valeur de $p_{x,y}$. Nous nous fondons sur la racine carrée de l'erreur quadratique moyenne et sur le taux de couverture pour comparer les estimateurs.

À la figure 1, nous notons que la RMSE diminue considérablement quand $p_{x,y}$ augmente, ce qui peut s'expliquer par la diminution de la dispersion de la variable dépendante conditionnellement à la variable indépendante à mesure que la corrélation entre les deux augmente. Nous constatons aussi que l'étalement de la RMSE est plus faible pour le grand domaine que pour le petit. Le classement des estimateurs du moins bon au meilleur en fonction de la RMSE est le suivant : i) quotient HT (RAT HT), ii) quotient Hajék (RAT HA), iii) quotient de rechange HT (ALTR HT), iv) quotient de rechange Hajék (ALTR HA) et v) quotient

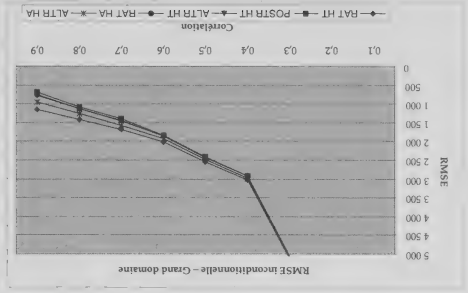
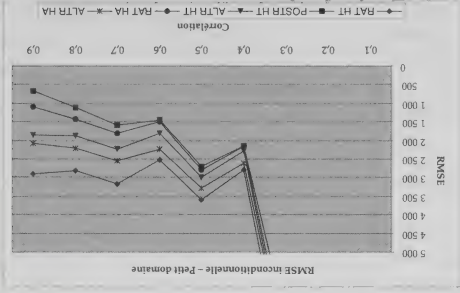


Figure 1. RMSE inconditionnelle



Les valeurs du biais relatif et du taux de couverture conditionnels des estimateurs sont résumés aux figures 3, 4a et 4b, pour la taille d'échantillon réalisée n_d pour le grand et pour le petit domaine, et pour deux valeurs de corrélation ($p_{x,y} = 0.90$ et $p_{x,y} = 0.60$).

Le tableau 5 résume les valeurs du biais conditionnel des estimateurs par quotient des estimateurs (2.1) à (2.4) et (2.6). Nous les avons obtenues à partir du tableau 3 en utilisant une variable auxiliaire unique.

Le tableau 5 résume les valeurs du biais conditionnel des estimateurs par quotient des estimateurs (2.1) à (2.4) et (2.6). Nous les avons obtenues à partir du tableau 3 en utilisant une variable auxiliaire unique.

$$V_2^d = \frac{M^d - 1}{M_d} \sum_{m=1}^{m=1} (\hat{Y}_{(m)}^d - \bar{Y}^d)^2$$

où

$$\bar{Y}^d = \frac{1}{M_d} \sum_{m=1}^{m=1} \hat{Y}_{(m)}^d$$

comme suit :

i. $BRM_d = (100/M_d) (\sum_{m=1}^{m=1} (\hat{Y}_{(m)}^d - Y_d^d) / Y_d^d)$, où M_d est le nombre d'échantillons de taille n_d ;

ii. le taux de couverture conditionnel est défini de la même façon que son analogue inconditionnel. La variance associée est

4.2 Résultats conditionnels

Nous avons étudié les propriétés conditionnelles des estimateurs au moyen i) du biais relatif moyen (BRM) conditionnel et ii) du taux de couverture conditionnel, définis comme suit :

Nous avons étudié les propriétés conditionnelles des estimateurs au moyen i) du biais relatif moyen (BRM) conditionnel et ii) du taux de couverture conditionnel, définis comme suit :

Nous avons étudié les propriétés conditionnelles des estimateurs au moyen i) du biais relatif moyen (BRM) conditionnel et ii) du taux de couverture conditionnel, définis comme suit :

L'équation qui précède donne le terme constant $\sigma^2 = \beta_2 b \left(\frac{1}{I} \frac{p_{x,y}^2}{\sigma^2} - 1 \right)$

de la variance de l'erreur. Nous avons utilisé pour les deux domaines des valeurs de corrélation $\rho_{x,y}$, courantes, variant de 0,1 à 0,9 par incrément de 0,1, qui nous ont donné neuf échantillons aléatoires ($M = 10\,000$) de taille 250 répétées à partir des populations. Pour chaque échantillon, nous avons estimé les totaux de domaine au moyen des estimateurs donnés au tableau 4. Nous n'incluons pas l'estimateur stratifié à posteriori de Häjek, $\hat{Y}_{d,lr}$, car il correspond exactement à son analogue d'Horvitz-Thompson, $\hat{Y}_{d,lr}$.

$$p_{x,y} = \frac{\sigma^2 b}{\sigma^2 b + \beta_2 b^2}.$$

900 et un petit domaine de taille 100. Nous avons généré les observations (y, x) dans chaque domaine en émettant l'hypothèse d'un modèle de quotient $y_k = \beta x_k + \varepsilon_k$ où $E(\varepsilon_k) = 0$ et $V(\varepsilon_k) = \sigma^2 x_k$. La valeur du coefficient β est de 1,0 et de 3,0, pour le grand et pour le petit domaine, respectivement. Nous avons généré la variable auxiliaire x au moyen d'une loi gamma $\Gamma(a, b)$, où $a = 3$ et $b = 16$. Nous avons également généré la variable dépendante y au moyen d'une loi gamma $\Gamma(A, B)$ telle que les paramètres A et B satisfassent $E(y_k) = \beta x_k = AB$ et $V(y_k) = \sigma^2 x_k = AB^2$. Le calcul de A et de B par résolution du système d'équations nous donne $A = \beta^2 / \sigma^2$ et $B = \sigma^2 / \beta$. Nous avons choisi le terme σ^2 de façon à satisfaire une corrélation fixée entre x et y définie par

Tableau 3
Biais et variances conditionnels des estimateurs (2.1) à (2.6)

Estimateur	Biais conditionnel	Variance conditionnelle estimée
$\hat{Y}_{d,lr}$	$N((w_d - w_d)/(1 - w_d))(\bar{y}_{U_d} - \bar{x}_{U_d})(\mathbf{B}^{ld})$	$N^2 \left[w_d^2/n_d \left((1 - f_d) v_{\varepsilon_{U_d}} + (w_d^2/n_d^2)(1 - f_d) v_{\varepsilon_{U_d}} \right) \right]$
$\hat{Y}_{d,lr}$	Presque 0	$(N^2/n_d) \left(\sum_{s_d} (a_{dk} e_{dk} - a_{dk})^2 / (n_d - 1) \right)$
$\hat{Y}_{d,lr}$	$N(w_d - w_d)(\bar{y}_{U_d} - \bar{x}_{U_d})(\mathbf{B}^3)$	$(N^2/n_d) \left(\sum_{s_d} (a_{dk} e_{dk} - a_{dk})^2 / (n_d - 1) \right)$
$\hat{Y}_{d,lr}$	$N(w_d - w_d)(\bar{y}_{U_d} - \bar{x}_{U_d})(\mathbf{B}^{ld})$	$(N^2/n_d) \left(\sum_{s_d} (a_{dk} e_{dk} - a_{dk})^2 / (n_d - 1) \right)$
$\hat{Y}_{d,lr}$	Presque 0	$(N^2/n_d) \left(\sum_{s_d} (a_{dk} e_{dk} - a_{dk})^2 / (n_d - 1) \right)$
$\hat{Y}_{d,lr}$	Presque 0	$(N^2/n_d) \left(\sum_{s_d} (a_{dk} e_{dk} - a_{dk})^2 / (n_d - 1) \right)$
$\hat{Y}_{d,lr}$	Presque 0	$(N^2/n_d) \left(\sum_{s_d} (a_{dk} e_{dk} - a_{dk})^2 / (n_d - 1) \right)$

Tableau 4
Estimateurs et termes d'erreur connexes

Estimateur	Version du quotient	Terme d'erreur
Quotient HT : $\hat{Y}_{d,lr}$	$\hat{Y}_{d,RAT} = \hat{Y}_{d,HT} \left(X / \hat{X}_{HT} \right)$	$e_{dk} = \hat{Y}_{d,RAT} - \hat{Y}_{d,HT} X_k$
Quotient stratifié à posteriori HT : $\hat{Y}_{d,lr}$	$\hat{Y}_{d,POSTR} = \hat{Y}_{d,HT} \left(X_d / \hat{X}_{d,HT} \right)$	$e_{dk} = \hat{Y}_{d,POSTR} - \hat{Y}_{d,HT} X_{dk}$
Quotient de rechange HT : $\hat{Y}_{d,lr}$	$\hat{Y}_{d,ALTR} = \hat{Y}_{d,HT} + (X_d - \hat{X}_{d,HT}) \left(\hat{Y}_{d,HT} / \hat{X}_{HT} \right)$	$e_{dk} = \hat{Y}_{d,ALTR} - \hat{Y}_{d,HT} X_{dk}$
Quotient de rechange de Häjek : $\hat{Y}_{d,lr}$	$\hat{Y}_{d,RAT} = \hat{Y}_{d,HA} + (X - \hat{X}_{HA}) \left(\hat{Y}_{d,HA} / \hat{X}_{HT} \right)$	$e_{dk} = \hat{Y}_{d,RAT} - \hat{Y}_{d,HA} X_k$
Quotient de rechange de Häjek : $\hat{Y}_{d,lr}$	$\hat{Y}_{d,ALTR} = \hat{Y}_{d,HA} - \hat{X}_{d,HA} \left(\hat{Y}_{d,HA} / \hat{X}_{HT} \right)$	$e_{dk} = \hat{Y}_{d,ALTR} - \hat{Y}_{d,HA} X_{dk}$

$$\mathbf{B}^{1d}_* = \left[E \left(\sum_s \mathbf{x}_k \mathbf{x}_k' \right) \right]^{-1} \left[E \left(\sum_s \mathbf{x}_k \mathbf{y}_k \right) \right] \times$$

$$(3.2) \quad \left[E \left(\sum_s \mathbf{x}_k \mathbf{y}_k \right) \right] \left[E \left(\sum_s \mathbf{x}_k \mathbf{x}_k' \right) \right]^{-1} \left[E \left(\sum_s \mathbf{y}_k \right) \right]$$

Le vecteur de régression estime \mathbf{B}^{1d} convergera vers \mathbf{B}^{1d}_* (sous les conditions appropriées) en probabilité d'inclusion du plan de sondage conditionnelle quand n_d et N_d augmentent.

Nous avons

$$\left[E \left(\sum_s \mathbf{x}_k \mathbf{x}_k' \right) \right]^{-1} \left[E \left(\sum_s \mathbf{x}_k \mathbf{y}_k \right) \right] = \sum_u \mathbf{x}_k \mathbf{x}_k' / N_c + \mathbf{R}_c$$

$$\left[E \left(\sum_s \mathbf{x}_k \mathbf{y}_k \right) \right]^{-1} \left[E \left(\sum_s \mathbf{y}_k \right) \right] = \sum_u \mathbf{x}_k \mathbf{y}_k / N_c + \mathbf{r}_c,$$

et

$$\mathbf{R}_c = \frac{1}{N} \sum_u \mathbf{x}_k \mathbf{x}_k' \left(\frac{1}{N} \sum_u \mathbf{y}_k \right) - \frac{1}{N} \sum_u \mathbf{x}_k \mathbf{y}_k \left(\frac{1}{N} \sum_u \mathbf{y}_k \right)$$

et

$$\mathbf{r}_c = \frac{1}{N} \sum_u \mathbf{x}_k \mathbf{y}_k \left(\frac{1}{N} \sum_u \mathbf{y}_k \right) - \frac{1}{N} \sum_u \mathbf{x}_k \mathbf{y}_k \left(\frac{1}{N} \sum_u \mathbf{y}_k \right) = 0.$$

Conséquemment, en utilisant le résultat 3.3 et en supposant que $(w_d - w^d)/(1 - w^d) = 0$, nous obtenons

$$\mathbf{B}^{1d}_* = \mathbf{B}^{1d}.$$

Définissons le « résidu conditionnel » pour la k^e unité comme étant

$$E^{dk}_* = y^{dk} - \mathbf{x}_k' \mathbf{B}^{1d}_*.$$

L'écart de $Y_{d,ln}$ par rapport à la valeur réelle Y_d peut s'écrire sous la forme

$$(3.4) \quad Y_{d,ln} - Y_d = - \sum_u E^{dk}_* + \frac{1}{N} \sum_s E^{dk}_* - \Delta^{1d}_*.$$

où

$$\Delta^{1d}_* = \left(\frac{1}{N} \sum_u \mathbf{x}_k - \sum_u \mathbf{x}_k \right) (\mathbf{B}^{1d}_* - \mathbf{B}^{1d}).$$

Dans l'équation (3.4), Δ^{1d}_* est d'un ordre inférieur à $N/n \sum_s E^{dk}_*$. Pour le montrer, notons que

$$\left[E \left(\frac{1}{N} \sum_s \mathbf{x}_k - \sum_u \mathbf{x}_k \right) \right] \left[E \left(\frac{1}{N} \sum_s \mathbf{x}_k - \sum_u \mathbf{x}_k \right) \right]^{-1} \left[E \left(\frac{1}{N} \sum_s \mathbf{x}_k - \sum_u \mathbf{x}_k \right) \right] = N \left(\frac{1}{N} \sum_s \mathbf{x}_k - \sum_u \mathbf{x}_k \right) \left(\frac{1}{N} \sum_s \mathbf{x}_k - \sum_u \mathbf{x}_k \right)'$$

où $(w_d - w^d)/(1 - w^d)$ devrait s'approcher de zéro.

En outre, comme nous l'avons mentionné plus haut, $\mathbf{B}^{1d} - \mathbf{B}^{1d}_*$ est proche du vecteur $\mathbf{0}$ en probabilité d'inclusion du plan de sondage conditionnel. Donc, $E^{dk}_* = y^{dk} - \mathbf{x}_k' \mathbf{B}^{1d}_* \doteq y^{dk} - \mathbf{x}_k' \mathbf{B}^{1d} = E^{dk}$. Par conséquent, nous pouvons écrire (3.4) sous la forme

4. ÉTUDE EN SIMULATION

Nous avons réalisé une étude en simulation pour illustrer les propriétés conditionnelles et inconditionnelles de la version par quotient des estimateurs (2.1) à (2.6). Nous nous sommes servis d'une population de 1 000 observations bivariées (y, x) pour étudier ces propriétés. Nous avons obtenu cette population par concaténation de deux domaines de population générés, à savoir un grand domaine de taille

tableaux 1 et 2.

Le biais et les variances conditionnels des cinq autres estimateurs peuvent être calculés de la même façon. Le tableau 3 résume ces propriétés. Les facteurs d'ajustement e^{dk} et les termes résiduels e^{dk} sont donnés dans les

et

$$V_{e^{dk}} = \left(\frac{1}{N} \sum_s e^{dk} - \sum_u e^{dk} \right) \left(\frac{1}{N} \sum_s e^{dk} - \sum_u e^{dk} \right)'$$

$$V_{E^{dk}_*} = \left(\frac{1}{N} \sum_s E^{dk}_* - \sum_u E^{dk}_* \right) \left(\frac{1}{N} \sum_s E^{dk}_* - \sum_u E^{dk}_* \right)'$$

$$V_{Y_{d,ln}} = \left(\frac{1}{N} \sum_s Y_{d,ln} - \sum_u Y_{d,ln} \right) \left(\frac{1}{N} \sum_s Y_{d,ln} - \sum_u Y_{d,ln} \right)'$$

et

$$V_{Y_{d,ln} - Y_d} = \left(\frac{1}{N} \sum_s (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right) \left(\frac{1}{N} \sum_s (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right)'$$

utilisons le résultat 3.4, la variance de population conditionnelle de $Y_{d,ln}$ et sa valeur estimée sont, respectivement

Dans (3.5), le terme $\sum_u E^{dk}$ est constant. Si nous

$$(3.7) \quad \left[\frac{1}{N} \sum_u (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right] \left(\frac{1}{N} \sum_u (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right)'$$

$$= N \left(\frac{1}{N} \sum_u (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right) \left(\frac{1}{N} \sum_u (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right)'$$

$$E \left[\left(\frac{1}{N} \sum_u (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right) \left(\frac{1}{N} \sum_u (Y_{d,ln} - Y_d) - \sum_u (Y_{d,ln} - Y_d) \right)' \right]$$

ditonnelle (3.6) comme suit :

$$(3.6) \quad \tilde{E}_{u_d} = \sum_u E^{dk} / N_d \text{ et } \tilde{E}_{u_d} = \sum_u E^{dk} / N. \text{ Puisque } \tilde{Y}_{u_d} = w_d \tilde{Y}_{u_d}, \text{ nous pouvons réexprimer l'espérance con-}$$

mativement :

L'espérance conditionnelle de $Y_{d,ln} - Y_d$ est approxi-

$$(3.5) \quad Y_{d,ln} - Y_d = - \sum_u E^{dk} + \frac{1}{N} \sum_s E^{dk}.$$

du vecteur de données (y_k, x_k) sont positifs pour tout $k \in U$. Les variances de population de $\hat{Y}_{d, \text{RAT}}^k$ et $\hat{Y}_{d, \text{POSTR}}^k$ sont $V(\hat{Y}_{d, \text{RAT}}^k) = A \sum_{U_d} (y_k - B_{2d} x_k)^2$ et $V(\hat{Y}_{d, \text{POSTR}}^k) = A \sum_{U_d} (y_{dk} - B_{1d} x_k)^2$, où $B_{2d} = Y_d / X_d$ et $B_{1d} = Y_d / X_d$.

La différence $V(\hat{Y}_{d, \text{RAT}}^k) - V(\hat{Y}_{d, \text{POSTR}}^k)$ peut être réécrite sous la forme :

$$A \sum_{U_d} (B_{1d} - B_{2d})^2 x_k^2 + 2A(B_{1d} - B_{2d}) \sum_{U_d} (y_k - B_{2d} x_k) x_k + A \sum_{U_d} (y_{dk} - B_{1d} x_k)^2.$$

Puisque le deuxième terme de cette expression peut être positif, négatif ou nul, la différence $V(\hat{Y}_{d, \text{RAT}}^k) - V(\hat{Y}_{d, \text{POSTR}}^k)$ peut être négative.

3.2 Propriétés conditionnelles

Pour un échantillon donné, soit n_d la taille d'échantillon réalisée de s_d . Nous pouvons utiliser le résultat qui

(2.6).
Résultat 3.3 : Soit \mathbf{z}_k un vecteur arbitraire de dimension p , c'est-à-dire $\mathbf{z}_k = (z_{k1}, \dots, z_{kp})'$ et supposons que $n_d \geq 1$. L'espérance conditionnelle de $\bar{\mathbf{z}}_s = n^{-1} \sum_{s_d} \mathbf{z}_k$ sachant n_d peut s'écrire :

$$E(\bar{\mathbf{z}}_s | n_d) = \frac{1}{n} \left[f^d \sum_{U_d} \mathbf{z}_k + f^d \sum_{U_d} \mathbf{z}_k \right]$$

$$= \mathbf{z}^u + \frac{w_d - 1}{w_d} (\mathbf{z}^u - \mathbf{z}^u) \quad (3.1)$$

où $\mathbf{z}^u = N^{-1} \sum_{U_d} \mathbf{z}_k$, $\mathbf{z}^u = N^{-1} \sum_{U_d} \mathbf{z}_k$, $w_d = n_d / n$, $f^d = n_d / N$, $f^d = n_d / N$, $f^d = n_d / N$ et $n_d = n - n_d$.

Preuve : Si nous réécrivons $\bar{\mathbf{z}}_s$ sous la forme

$$\frac{1}{n} \left(\sum_{s_d} \mathbf{z}_k + \sum_{s_d} \mathbf{z}_k \right),$$

nous obtenons

$$E(\bar{\mathbf{z}}_s | n_d) = \frac{1}{n} \left[\frac{n_d}{N} \sum_{U_d} \mathbf{z}_k + \frac{n - n_d}{N} \sum_{U_d} \mathbf{z}_k \right]$$

où $s_d = \{k \in s \text{ et } k \notin s_d\}$ et

$$U_d = \{k \in U \text{ et } k \notin U_d\}.$$

Puisque $\sum_{U_d} \mathbf{z}_k = \sum_{U_d} \mathbf{z}_k - \sum_{U_d} \mathbf{z}_k$, nous obtenons le résultat requis, c'est-à-dire

$$E(\bar{\mathbf{z}}_s | n_d) = \mathbf{z}^u + \frac{w_d - 1}{w_d} (\mathbf{z}^u - \mathbf{z}^u).$$

Résultat 3.4 : La variance de population conditionnelle de $\bar{\mathbf{z}}_s$ sachant n_d peut s'écrire

$$V(\bar{\mathbf{z}}_s | n_d) = \frac{w_d}{w_d^2} (1 - f^d) \mathbf{V}^u + \frac{n_d}{w_d^2} (1 - f^d) \mathbf{V}^{z_{s_d}},$$

$$\mathbf{V}^u = \frac{1}{n} \sum_{U_d} (\mathbf{z}_k - \mathbf{z}^u) (\mathbf{z}_k - \mathbf{z}^u)',$$

$$\mathbf{V}^{z_{s_d}} = \frac{1}{n} \sum_{s_d} (\mathbf{z}_k - \mathbf{z}^{z_{s_d}}) (\mathbf{z}_k - \mathbf{z}^{z_{s_d}})',$$

$$\mathbf{V}(\bar{\mathbf{z}}_s | n_d) \text{ est donné par}$$

$$\mathbf{V}(\bar{\mathbf{z}}_s | n_d) = N_{-1}^{-1} \sum_{U_d} \mathbf{z}_k \text{ et } w_d = 1 - w_d.$$

L'estimateur de la variance de population conditionnelle avec $\bar{\mathbf{z}}^u = N_{-1}^{-1} \sum_{U_d} \mathbf{z}_k$ et $w_d = 1 - w_d$ avec $\bar{\mathbf{z}}^{z_{s_d}} = N_{-1}^{-1} \sum_{s_d} \mathbf{z}_k$.

où

Preuve : Elle découle de l'emploi d'arguments semblables à ceux utilisés pour le résultat 3.3. Nous commençons par illustrer comment on peut utiliser le résultat 3.3 pour obtenir le biais conditionnel pour les estimateurs plus simples des totaux de domaines. Ces estimateurs incluent l'estimateur d'Horvitz-Thompson $\hat{Y}_{d, \text{HT}}^k$, ainsi que l'estimateur par quotient stratifié à posteriori $\hat{Y}_{d, \text{POSTR}}^k = (X_d / \hat{X}_{d, \text{HT}}) \hat{Y}_{d, \text{HT}}^k$. Soit \mathbf{z}_k la variable de domaine y_{dk} . En utilisant le résultat 3.3, nous avons que $E(\hat{Y}_{d, \text{HT}}^k | n_d) = N w_d \hat{Y}_{d, \text{HT}}^k$, où $\hat{Y}_{d, \text{HT}}^k = Y_d / N_d$. Le biais conditionnel de $\hat{Y}_{d, \text{HT}}^k$ sachant n_d est par conséquent $\text{Biais}(\hat{Y}_{d, \text{HT}}^k | n_d) = N(w_d - 1) \hat{Y}_{d, \text{HT}}^k$. Pour l'estimateur par quotient stratifié à posteriori, notons que $\hat{Y}_{d, \text{POSTR}}^k = \hat{Y}_{d, \text{HT}}^k - Y_d^d / (Y_d^d / X_d^d) \hat{X}_{d, \text{HT}}^k$. Si nous définissons \mathbf{z}_k comme étant $y_{dk} - (Y_d^d / X_d^d) x_{dk}$, nous obtenons que $\text{Biais}(\hat{Y}_{d, \text{POSTR}}^k | n_d) = 0$.

Passons maintenant à l'évaluation du biais et de la variance conditionnels des estimateurs (2.1) à (2.6). Nous illustrons la procédure que pour l'estimateur par régression $\hat{Y}_{d, \text{R}}^k$, puisque les étapes sont les mêmes pour les autres estimateurs. Conditionnellement à n_d , la distribution de s_d est celle d'un EASSR. Autrement dit, pour chaque échantillon s_d , n_d peut être considéré comme ayant été sélectionné à partir de N_d . Nous exprimons $\hat{Y}_{d, \text{R}}^k$ sous la forme $\hat{Y}_{d, \text{R}}^k = \sum_{U_d} y_k + N/n \sum_{s_d} e_{dk}$, où $e_{dk} = y_{dk} - x_k^d B_{1d}$ et $\hat{Y}_{d, \text{R}}^k = \sum_{U_d} y_k + N/n \sum_{s_d} e_{dk}$.

Comme l'ont fait Särndal et Hidiroglou (1989), nous définissons le vecteur de régression conditionnel \mathbf{B}_{1d}^k comme suit

Remarque 3.2 : (Calage sur des données auxiliaires de domaine). Estévaou et coll. (1999) ont discuté de certains estimateurs représentés dans les tableaux 1 et 2 pour le cas d'une variable auxiliaire unique x_k . Ils ont obtenu leurs estimateurs en tenant compte de l'information de domaine au moyen de variables auxiliaires et (ou) de totaux de contrôle.

Dans la suite, nous supposons que l'échantillon s de taille n a été sélectionné par échantillonnage aléatoire simple sans remise (EASSR) à partir d'un univers de taille N . L'estimation de la variance inconditionnelle des estimateurs de type Horvitz-Thompson et de type Hajék pour ce plan d'échantillonnage est :

$$V(\hat{Y}^{d, \ell_{Tj}}) = \sum_s \frac{n}{N^2(1-f)} \frac{n-1}{\sum_s (a_{dk} e^{dk} - a_d e^{\ell})^2} \quad (2.12)$$

où $\overline{a_d e^{\ell}} = \sum_s (a_{dk} e^{dk} / n)$ et $f = n / N$ est la fraction d'échantillonnage.

3.1 Propriétés inconditionnelles

Le choix entre les divers estimateurs par régression devrait se fonder sur le niveau auquel les totaux auxiliaires sont disponibles, ainsi que sur le biais et la variance. Tous les estimateurs susmentionnés sont asymptotiquement inconditionnellement sans biais; toutefois, leur variance diffère. Nous comparons les variances de population inconditionnelles des six estimateurs de domaine par régression (2.1) à (2.6) en faisant la distinction entre deux cas : i) inclusion d'un terme de coordonnée à l'origine dans la régression et (ii) pas d'inclusion de terme de coordonnée à l'origine dans la régression.

Résultat 3.2 : Supposons qu'une coordonnée à l'origine est incluse dans la régression, que $c_k = c$ pour tout $k \in U$ et que $N > p$, où p représente le nombre de variables auxiliaires. Les inégalités qui suivent sont vérifiées pour les variances de population des estimateurs de domaine par régression (2.1) à (2.6) :

$$\begin{aligned} \text{i) } & V(\hat{Y}^{d, \ell_{Tj}}) > V(\hat{Y}^{d, \ell_{Tj}}); \quad V(\hat{Y}^{d, \ell_{Tj}}) \leq V(\hat{Y}^{d, \ell_{Tj}}); \quad V(\hat{Y}^{d, \ell_{Tj}}) \\ & \text{peut être inférieure, égale ou supérieure à } V(\hat{Y}^{d, \ell_{Tj}}). \\ \text{ii) } & V(\hat{Y}^{d, \ell_{Tj}}) > V(\hat{Y}^{d, \ell_{Tj}}) \text{ et } V(\hat{Y}^{d, \ell_{Tj}}) > V(\hat{Y}^{d, \ell_{Tj}}); \quad V(\hat{Y}^{d, \ell_{Tj}}) \\ & \text{peut être inférieure, égale ou supérieure à } V(\hat{Y}^{d, \ell_{Tj}}). \end{aligned}$$

Preuve : Dans le cas de l'échantillonnage aléatoire simple sans remise, $V(\hat{Y}^{d, \ell_{Tj}}) = A \sum_U (E^{dk} - \bar{E}^U)^2$ pour $\ell = 1$, $2, 3$, où $A = N^2(1-f) / (n(N-1))$ et $\bar{E}^U = \sum_U E^{dk} / N$. Sachant que la régression contient une coordonnée à l'origine, il s'ensuit que $\sum_U E^{dk} = 0$ ou que $\sum_U E^{dk} = 0$, selon l'estimateur par régression utilisé. Nous

Preuve : Nous montrons d'abord comment arriver à la forme de l'estimateur par régression. Si nous définissons le vecteur de données auxiliaires \mathbf{z}_k comme étant $\mathbf{z}_k' = (x_{0k}, \mathbf{x}_k')$, l'estimateur par régression est

$$\hat{Y}^{\ell_T} = \hat{Y}^{\text{HT}} + (\mathbf{Z} - \hat{\mathbf{Z}}^{\text{HT}})' \hat{\mathbf{B}}_z$$

$$\hat{\mathbf{B}}_z = \left(\sum_s \mathbf{w}_k \mathbf{z}_k \mathbf{z}_k' / c_k \right)^{-1} \left(\sum_s \mathbf{w}_k \mathbf{z}_k Y_k / c_k \right),$$

Si $\mathbf{Y}^{x_0} = \mathbf{1}$, $\hat{\mathbf{Y}}^{\ell_T}$ est exactement équivalent à $\hat{Y}^{\ell_T} = \mathbf{Z}' \hat{\mathbf{B}}_z$. Si nous décomposons $\hat{\mathbf{B}}_z$ en

$$\hat{\mathbf{B}}_z' = (\hat{B}_0, \hat{\mathbf{B}}_x)',$$

nous obtenons $\hat{Y}^{\ell_T} = N \hat{B}_0 + \sum_U \mathbf{x}_k' \hat{\mathbf{B}}_x$, où $\hat{B}_0 = \bar{y}_s$ et

$$\hat{\mathbf{B}}_x = \left(\frac{\sum_s \mathbf{w}_k (\mathbf{x}_k - \bar{\mathbf{x}}_s)(\mathbf{x}_k - \bar{\mathbf{x}}_s)'}{c_k} \right)^{-1} \times \frac{\sum_s \mathbf{w}_k (\mathbf{x}_k - \bar{\mathbf{x}}_s)(Y_k - \bar{y}_s)}{c_k}.$$

Donc, la forme de Hajék de l'estimateur par régression est

$$\hat{Y}^{\ell_T} = \hat{Y}^{\text{HA}} + (\mathbf{X}^U - \hat{\mathbf{X}}^{\text{HA}})' \hat{\mathbf{B}}_x.$$

La régression de y_k sur

$$(\mathbf{x}_k')' = (1, (\mathbf{x}_k - \bar{\mathbf{x}}^U)')$$

donne le vecteur de régression estimé $\hat{\mathbf{B}}^* = (\hat{B}_0^*, \hat{\mathbf{B}}_x^*)'$, où

$$\hat{\mathbf{B}}_x^* = \left(\frac{c_k}{\sum_s \mathbf{w}_k (\mathbf{x}_k - \bar{\mathbf{x}}_s)(\mathbf{x}_k - \bar{\mathbf{x}}_s)'} \right)^{-1} \frac{c_k}{\sum_s \mathbf{w}_k (\mathbf{x}_k - \bar{\mathbf{x}}_s)(Y_k - \bar{y}_s)}.$$

et $\hat{B}_0^* = \bar{y}_s + (\bar{\mathbf{x}}^U - \bar{\mathbf{x}}_s)' \hat{\mathbf{B}}_x^*$. En substituant \hat{B}_1^* dans $\hat{Y}^{\ell_T} = N \hat{B}_0^* + \sum_U \mathbf{x}_k' \hat{\mathbf{B}}_x^*$, nous obtenons la forme de Hajék \hat{Y}^{ℓ_T} .

Remarque 3.1 : (Additivité). Supposons que les domaines U_d soient mutuellement exclusifs ($U_{d_1} \cap U_{d_2} = \emptyset$ pour $d_1 \neq d_2$) et exhaustifs ($\bigcup_{d=1}^D U_d = U$). L'additivité sur de tels domaines signifie que $\sum_{d=1}^D \hat{Y}^{d, \ell_{Tj}} = \hat{Y}^{\ell_T}$ où

$$\hat{Y}^{\ell_T} = \hat{Y}^{\text{HT}} + (\mathbf{X} - \hat{\mathbf{X}}^{\text{HT}})' \hat{\mathbf{B}}.$$

La propriété d'additivité de $\hat{Y}^{d, \ell_{Tj}}$ est désirable, car il est possible d'utiliser un ensemble unique de poids de calage, $w_k^{d, \ell_{Tj}}$, de façon répétée pour produire des estimations de domaine ponctuelles. Seuls deux des six estimateurs, $\hat{Y}^{d, \ell_{Tj}}$ et $\hat{Y}^{d, \ell_{Tj}}$, sont additifs sur tous les domaines de ce genre.

$$V(\tilde{Y}_{d,lr_j}) = \left\{ \sum \sum^U \Delta_{k\ell} \left(\frac{E_{dk} - E_{U_d}}{E_{d\ell} - \tilde{E}_{U_d}} \right) \left(\frac{\pi_k}{E_{d\ell} - \tilde{E}_{U_d}} \right) \right\} \text{ pour } j = 2, 3 \quad (2.10)$$

$$V(\tilde{Y}_{d,lr_j}) = \sum \sum^s \Delta_{k\ell} \left(\frac{a_{k\ell} e_{dk}}{a_{d\ell} e_{d\ell}} \right) \left(\frac{\pi_k}{a_{d\ell} e_{d\ell}} \right) \quad \text{pour } j = 1, 2, 3 \quad (2.11)$$

matrices de type Hájek \tilde{Y}_{d,lr_j} ($j = 1, 2, 3$) sont :

$$V(\tilde{Y}_{d,lr_j}) = \sum \sum^s \Delta_{k\ell} \left(\frac{a_{k\ell} e_{dk}}{a_{d\ell} e_{d\ell}} \right) \left(\frac{\pi_k}{a_{d\ell} e_{d\ell}} \right) \quad (2.9)$$

où $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$; $\pi_{k\ell} = \Pr \{ k, \ell \in s \}$ avec les E_{dk} , e_{dk} , a_{dk} appropriés tels que définis au tableau 1.

La variance de population inconditionnelle approximative et la variance estimée correspondante des estimateurs de type Hájek

Tableau 1 Facteurs d'ajustement et résidus pour les estimateurs par régression d'Horvitz-Thompson

Estimateur	Dependant du domaine	Facteur d'ajustement : a_{dk}	Résidus
\tilde{Y}_{d,lr_1}	Non	$1 + (\mathbf{X} - \hat{\mathbf{X}}^{\text{HT}})' \left(\sum^s \frac{C_k}{w_k \mathbf{x}_k \mathbf{x}_k'} \right)^{-1} \frac{C_k}{\mathbf{x}_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$
\tilde{Y}_{d,lr_2}	Oui	$\delta_{dk} \left(1 + (\mathbf{X}^d - \hat{\mathbf{X}}_{d,\text{HT}})' \left(\sum^s \frac{C_k}{w_k \mathbf{x}_k \mathbf{x}_k'} \right)^{-1} \frac{C_k}{\mathbf{x}_k} \right)$	$E_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{2d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{2d}$
\tilde{Y}_{d,lr_3}	Oui	$\delta_{dk} + (\mathbf{X}^d - \hat{\mathbf{X}}_{d,\text{HT}})' \left(\sum^s \frac{C_k}{w_k \mathbf{x}_k \mathbf{x}_k'} \right)^{-1} \frac{C_k}{\mathbf{x}_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_3$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_3$

Tableau 2 Facteurs d'ajustement et résidus pour les estimateurs de type Hájek

Estimateur	Dependant du domaine	Facteur d'ajustement : a_{dk}	Résidus
\tilde{Y}_{d,lr_1}	Oui	$\frac{N_d^d}{N^d} + (\mathbf{X} - \hat{\mathbf{X}}^{\text{HA}})' \left(\sum^s \frac{C_k}{w_k \mathbf{x}_k \mathbf{x}_k'} \right)^{-1} \frac{C_k}{\mathbf{x}_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{1d}$
\tilde{Y}_{d,lr_2}	Oui	$\frac{N_d^d}{N^d} + (\mathbf{X}^d - \hat{\mathbf{X}}_{d,\text{HA}})' \left(\sum^s \frac{C_k}{w_k \mathbf{x}_k \mathbf{x}_k'} \right)^{-1} \frac{C_k}{\mathbf{x}_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{2d}$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_{2d}$
\tilde{Y}_{d,lr_3}	Oui	$\frac{N_d^d}{N^d} + (\mathbf{X}^d - \hat{\mathbf{X}}_{d,\text{HA}})' \left(\sum^s \frac{C_k}{w_k \mathbf{x}_k \mathbf{x}_k'} \right)^{-1} \frac{C_k}{\mathbf{x}_k}$	$E_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_3$ $e_{dk} = y_{dk} - \mathbf{x}_k' \hat{\mathbf{B}}_3$

la forme $\tilde{Y}_{lr} = Y_{\text{HA}}'(\mathbf{B}_x - \hat{\mathbf{X}}^{\text{HA}})' \hat{\mathbf{B}}_x$ de Hájek. Nous pouvons obtenir les divers estimateurs de domaine par régression de y_{dk} sur

L'estimateur par régression du total $\hat{Y}_{lr} = N \hat{B}_1'$ est égal à

$$\hat{B}_1' = \hat{y}_s' + (\hat{\mathbf{x}}^U - \hat{\mathbf{x}}_s)' \hat{\mathbf{B}}_x, \text{ avec } \hat{y}_s = \hat{Y}_{lr}^{\text{HT}} / N \text{ et } \hat{\mathbf{x}}_s = \sum^s (w_k (\mathbf{x}_k - \hat{\mathbf{x}}_s) / C_k)$$

où $\hat{\mathbf{B}}_x = \left(\left(\sum^s w_k (\mathbf{x}_k - \hat{\mathbf{x}}_s) (\mathbf{x}_k - \hat{\mathbf{x}}_s)' / C_k \right)^{-1} \right) \times \left(\sum^s w_k (\mathbf{x}_k - \hat{\mathbf{x}}_s) / C_k \right)$,
 $\hat{\mathbf{B}}_0' = (\hat{B}_0', \hat{\mathbf{B}}_x')$,

où $\hat{\mathbf{x}}^U = N^{-1} \sum^U \mathbf{x}_k$. Le vecteur de régression résultant est

régression de y_k sur

Résultat 3.1 : Nous pouvons obtenir l'estimateur par régression de type Hájek comme un sous-produit de la

pour les estimateurs de type Horvitz-Thompson que pour ceux de type Hájek.

forme de la variance inconditionnelle estimée est la même sont définis au tableau 2. Il convient de souligner que la

Cas 2

Le paramètre de régression de population

$$\mathbf{B}_{2d} = \left(\sum_{i_d} w_k \mathbf{x}_k' / c_k \right)^{-1} \sum_{i_d} w_k \mathbf{x}_k y_k / c_k$$

est estimé en calculant la régression de y_k sur \mathbf{x}_k pour chaque domaine U_d séparément. Son estimateur est donné

par

$$\hat{\mathbf{B}}_{2d} = \left(\sum_{i_d} w_k \mathbf{x}_k' / c_k \right)^{-1} \sum_{i_d} w_k \mathbf{x}_k y_k / c_k,$$

et l'estimateur de régression résultant d'un total de domaine

$$\hat{Y}_{d,lrj} = \hat{Y}_{d,HT} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HT})' \hat{\mathbf{B}}_{2d} \quad (2.2)$$

où $\hat{\mathbf{X}}_d = \sum_s w_k \mathbf{x}_{dk}$ avec \mathbf{x}_{dk} défini de la même façon que

Y_{dk} .
Exemple : L'estimateur d'Horvitz-Thompson stratifié à posteriori (noté POSTR) donné par $\hat{Y}_{d,POSTR} = X_d \hat{R}_{2d}$, où

$$\hat{R}_{2d} = \hat{Y}_{d,HT} / \hat{X}_{d,HT}.$$

Cas 3

Le paramètre de régression de population

$$\mathbf{B}_3 = \left(\sum_U (\mathbf{x}_k \mathbf{x}_k' / c_k) \right)^{-1} \sum_U (\mathbf{x}_k y_k / c_k)$$

est estimé en calculant la régression de y_k sur \mathbf{x}_k en utilisant toutes les unités comprises dans U . L'estimateur

correspondant est

$$\hat{\mathbf{B}}_3 = \left(\sum_s w_k \mathbf{x}_k' / c_k \right)^{-1} \sum_s w_k \mathbf{x}_k y_k / c_k,$$

ce qui donne l'estimateur par régression

$$\hat{Y}_{d,lrj} = \hat{Y}_{d,HT} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HT})' \hat{\mathbf{B}}_3. \quad (2.3)$$

Exemple : L'estimateur par quotient de rechange (noté ALTR pour *alternative*) donné par $\hat{Y}_{d,ALTR} = \hat{Y}_{d,HT} +$

$$(\mathbf{X}_d - \hat{\mathbf{X}}_{d,HT})' \hat{\mathbf{R}}_3, \text{ où } \hat{\mathbf{R}}_3 = \hat{Y}_{d,HT} / \hat{X}_{d,HT}.$$

2.2 Estimateurs de type Hájek

Les estimateurs (2.1) à (2.3) appartiennent à la famille des estimateurs d'Horvitz-Thompson. Si nous intégrons également la taille connue du domaine de population N_d dans l'estimation, nous obtenons les versions « Hájek » des estimateurs par régression d'Horvitz-Thompson définis à la

section précédente. Nous obtenons les estimateurs par régression de Hájek en remplaçant $\hat{Y}_{d,HT}$, $\hat{\mathbf{X}}_{d,HT}$ et $\hat{\mathbf{X}}_{HT}$

par

$$\hat{Y}_{d,HA} = \left(N_d / \hat{N}_d \right) \hat{Y}_{d,HT}, \quad \hat{\mathbf{X}}_{d,HA} = \left(N_d / \hat{N}_d \right) \hat{\mathbf{X}}_{d,HT}.$$

et

et

$$\hat{\mathbf{X}}_{HA} = \left(N / \hat{N} \right) \hat{\mathbf{X}}_{HT},$$

où $\hat{N}_d = \sum_s w_k$ et $\hat{N} = \sum_s w_k$. Les estimateurs sont presque conditionnellement sans biais pour un n_d donné, tandis que leurs analogues d'Horvitz-Thompson n'ont pas cette propriété. Les « \mathbf{B} » contenus dans les estimateurs par régression de Hájek correspondent exactement à leurs analogues d'Horvitz-Thompson.

Cas 4

$$\hat{Y}_{d,lrj} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{HA})' \hat{\mathbf{B}}_{1d}. \quad (2.4)$$

Exemple : L'estimateur par quotient de Hájek donné par $\hat{Y}_{d,RAT} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{HA})' \hat{\mathbf{R}}_{1d}$.

Cas 5

$$\hat{Y}_{d,lrj} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})' \hat{\mathbf{B}}_{2d}. \quad (2.5)$$

Exemple : L'estimateur par quotient stratifié à posteriori de Hájek donné par $\hat{Y}_{d,POSTR} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})' \hat{\mathbf{R}}_{2d}$.

Cet estimateur est identique à l'estimateur stratifié à posteriori d'Horvitz-Thompson.

Cas 6

$$\hat{Y}_{d,lrj} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})' \hat{\mathbf{B}}_3. \quad (2.6)$$

Exemple : L'estimateur par quotient de rechange de Hájek donné par $\hat{Y}_{d,ALTR} = \hat{Y}_{d,HA} + (\mathbf{X}_d - \hat{\mathbf{X}}_{d,HA})' \hat{\mathbf{R}}_3$.

3. PROPRIÉTÉS DES ESTIMATEURS DE DOMAINE

Nous pouvons exprimer les estimateurs (2.1) à (2.6) sous

la forme :

$$\hat{Y}_{d,lrj} = \sum_s w_k a_{dk} y_{dk} = \sum_s \tilde{w}_{dk} y_{dk} \quad (2.7)$$

où a_{dk} est un facteur d'ajustement qui peut dépendre ou non du domaine. Le produit du poids de sondage w_k et du facteur d'ajustement a_{dk} est appelé poids de régression (ou poids de calage) \tilde{w}_{dk} . Les tableaux 1 et 2 résument les valeurs de ces facteurs, ainsi que des résidus nécessaires pour l'estimation de la variance inconditionnelle. Les résidus de population et d'échantillon sont représentés par E_{dk} et e_{dk} respectivement. La variable indicatrice δ_{dk} est égale à l'unité si $k \in U_d$ et est nulle autrement.

La variance de population approximative et la variance estimée correspondante des estimateurs d'Horvitz-Thompson $\hat{Y}_{d,lrj}$ ($j = 1, 2, 3$) sont :

$$V(\hat{Y}_{d,lrj}) = \sum_U \sum_{k \in U} \Delta_{k \ell} \left(\frac{E_{dk}}{E} \right) \left(\frac{E_{dk}}{E} \right) \left(\frac{E_{dk}}{E} \right) \quad (2.8)$$

conditionnalisation est que les totaux de population sont connus pour chaque domaine. Dans le cas de l'échantillonnage aléatoire simple, on suppose que le nombre d'unités dans le domaine de population est connu.

L'objectif principal du présent article est d'étudier les propriétés inconditionnelles et conditionnelles d'un certain nombre d'estimateurs de totaux de domaine en présence de données auxiliaires dans le contexte de l'échantillonnage aléatoire simple sans remise (EASSR). Nous établirons ces propriétés conditionnelles en conditionnant sur des tailles d'échantillon fixes dans chaque domaine.

La présentation de l'article est la suivante. À la section 2, nous décrivons plusieurs estimateurs de totaux de domaine, puis, à la section 3, nous présentons leurs propriétés inconditionnelles et conditionnelles. À la section 4, nous présentons les résultats d'une étude en simulation pour le cas de l'estimateur par quotient des totaux de domaine. Enfin, à la section 5, nous tirons certaines conclusions.

2. ESTIMATEURS DE TOTAUX DE DOMAINE

Nous commençons par introduire la notation afin d'établir le cadre dans lequel nous évaluerons les propriétés des divers estimateurs des totaux de domaine. Soit $U = \{1, \dots, k, \dots, N\}$ la population finie. Nous tirons un échantillon « s » à partir de cette population conformément à un plan d'échantillonnage $P(s)$. Soient π_k et $\pi_{k'}$ les probabilités d'inclusion de premier et de deuxième ordre. Le total de domaine $X_d = \sum_{k \in U} x_k$ est le paramètre d'intérêt pour une variable « y ». Un domaine U_d ($d = 1, \dots, D$) est toute sous-population de U pour laquelle une estimation distincte peut être nécessaire, avant ou après le stade de la planification. Le nombre d'unités de population dans le domaine U_d est représenté par N_d , tel que $N = \sum_{d=1}^D N_d$ pour D domaines mutuellement exclusifs et épuisants couvrant l'ensemble de la population. L'échantillon s est subdivisé de façon correspondante en D domaines $s_1, \dots, s_d, \dots, s_D$ où $s_d = U_d \cap s$. La taille d'échantillon réalisée dans s_d est une variable aléatoire que nous représentons par n_d . Notons que la somme des n_d sur les domaines non chevauchants et épuisants de l'échantillon est égale à n . Un estimateur du total de domaine $X_d = \sum_{k \in U_d} x_k$ n'utilisant pas de données auxiliaires est donné par $\hat{X}_{d,HT} = \sum_{k \in U_d} w_k y_k = \sum_{k \in U_d} w_k^* y_{dk}$ où $w_k^* = \pi_k^{-1}$ et y_{dk} est égal à y_k si $k \in U_d$ et à 0, autrement.

L'information auxiliaire, sous la forme d'un vecteur \mathbf{x} de dimension p , peut exister à divers niveaux d'agrégation. On peut connaître le vecteur pour chaque unité de la population ou pour les sous-ensembles $U_g \subseteq U$ ($g = 1, \dots, G$) de la population U qui peuvent coïncider avec les domaines U_d . Nous représentons ces totaux connus par $\mathbf{X}_g = \sum_{k \in U_g} \mathbf{x}_k$; ils

$$\hat{Y}_{lr} = \hat{Y}_{HT} + \left(\mathbf{X} - \hat{\mathbf{X}}^{HT} \right) \mathbf{B}^g, \text{ où } \hat{\mathbf{X}}^{HT} = \sum_{k \in U} w_k^* \mathbf{x}_k. \quad (1.1)$$

Pour simplifier les choses, nous supposons que $g = 1$ (par exemple un groupe unique U), ce qui donne l'estimateur par régression simple $\hat{Y}_{lr} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}^{HT}) \mathbf{B}$, où $\hat{\mathbf{X}}^{HT} = \sum_{k \in U} w_k^* \mathbf{x}_k$.

Nous considérons six estimateurs pour estimer le total de population du domaine X_d , selon que nous utilisons les totaux de domaine \mathbf{X}_d ou le total de population \mathbf{X} , et que nous construisons l'estimation par régression au niveau du domaine ou de la population. Les estimateurs sont de type Horvitz-Thompson ou « Hájek ». Nous donnons un exemple de l'estimateur par quotient associé à chacun de ces estimateurs.

2.1 Estimateurs de type Horvitz-Thompson

Cas 1

Nous supposons que l'information auxiliaire \mathbf{x}_k est disponible au niveau de la population U , $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$ et qu'on calcule la régression des variables y_{dk} propres au domaine sur \mathbf{x}_k , $k \in U$. L'estimation du paramètre de régression de population résultant $\mathbf{B}^{ld} = (\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' / c_k)^{-1} (\sum_{k \in U} \mathbf{x}_k y_{dk} / c_k)$ est donnée par $\mathbf{B}^{ld} = (\sum_{k \in U} w_k^* \mathbf{x}_k \mathbf{x}_k' / c_k)^{-1} (\sum_{k \in U} w_k^* \mathbf{x}_k y_{dk} / c_k)$ et l'estimateur résultant du total de population X_d est

$$\hat{Y}_{d,lr} = \hat{Y}_{d,HT} + (\mathbf{X} - \hat{\mathbf{X}}^{HT}) \mathbf{B}^{ld}. \quad (2.1)$$

première fois par Hidiroglou (1991), est décrit plus en détail dans Estévez et coll. (1995).

Si l'on connaît les totaux des données auxiliaires au niveau du domaine, $\mathbf{X}_d = \sum_{k \in U_d} \mathbf{x}_k$, il est possible de construire deux estimateurs de X_d (cas 2 et 3), suivant la méthode d'estimation du paramètre de régression de population.

Estimation par domaine par la régression linéaire

MICHAEL A. HIDROGLOU et ZDENEK PATAK¹

RÉSUMÉ

L'un des objectifs principaux d'une enquête par sondage est d'estimer les moyennes et les totaux de domaines d'intérêt. Ces domaines sont déterminés avant que l'enquête soit réalisée (domaines primaires) ou après (domaines secondaires). La fiabilité des estimations connexes dépend de la variabilité de la taille de l'échantillon ainsi que des variables y d'intérêt. Il est impossible de tenir compte de cette variabilité en l'absence d'information auxiliaire sur des sous-groupes de la population. Toutefois, si l'on dispose de données auxiliaires, on peut contrôler dans une certaine mesure la fiabilité estimée des estimations résultantes. Dans le présent article, nous étudions les améliorations possibles de la fiabilité des estimations de domaine calculées en utilisant des données auxiliaires. Nous utilisons une approche conditionnelle pour comparer les propriétés (biais, couverture, efficacité) de divers estimateurs utilisant des données auxiliaires.

MOTS CLÉS : Estimation par domaine; données auxiliaires; propriétés conditionnelles.

1. INTRODUCTION

L'un des objectifs principaux d'une enquête par sondage est d'estimer les moyennes et les totaux d'un certain nombre de caractéristiques associées aux unités d'une population finale U . Les données sont souvent utilisées pour réaliser des études analytiques, comme la comparaison de moyennes et de totaux pour des sous-groupes de la population. Ces sous-groupes sont appelés *domaines d'intérêt*. Hartley (1959) a été l'un des premiers à essayer d'unifier la théorie de l'estimation par domaine. Il expose dans son article la théorie pour un certain nombre de plans d'échantillonnage pour lesquels l'estimation par domaine présentait un intérêt. Il y discute principalement des estimateurs n utilisant pas d'information auxiliaire. Toutefois, il considère le cas de l'estimateur par quotient où les totaux de population sont connus pour les domaines. L'utilisation de données auxiliaires dans le contexte de l'estimation par domaine a été abordée dans plusieurs articles. Sæmstad, Swensson et Wretman (1992) proposent un traitement unifié de l'estimation par domaine avec données auxiliaires. Estevao, Hidroglou et Sæmstad (1995) ont été les premiers à reconnaître que la pondération tenant compte des données auxiliaires pouvait ou non dépendre du domaine. Estevao et Sæmstad (1999) discutent des propriétés souhaitables des estimateurs par régression des totaux de domaine utilisant des données auxiliaires.

L'existence de données auxiliaires multivariées soulève un certain nombre de questions dans le contexte de l'estimation par domaine. En voici quelques-unes. Quelle est l'effet de l'existence de données auxiliaires pour le domaine d'intérêt si l'on ne les possède pas à l'échelle de la population? Comment calcule-t-on des estimations valides des totaux de domaine par régression des totaux de domaine utilisant des données auxiliaires? Comment calcule-t-on des estimations valides des totaux de domaine par régression des totaux de domaine utilisant des données auxiliaires? Comment calcule-t-on des estimations valides des totaux de domaine par régression des totaux de domaine utilisant des données auxiliaires?

Dans le cas de l'estimation par domaine, le nombre d'unités appartenant à un domaine particulier est une variable aléatoire. Dans ce contexte, les sous-ensembles reconnus-sables sont ceux pour lesquels la taille d'échantillon est fixée dans chaque domaine. On peut alors fonder la comparaison des propriétés statistiques conditionnelles (c'est-à-dire le biais et l'erreur quadratique moyenne) des divers estimateurs sur ces sous-ensembles. Le processus de

d'estimation de variance d'échantillonnage d'enquête. On notera avec intérêt que le simple estimateur de variance serait sans biais si tous les donneurs étaient extérieurs au domaine ($d_{ij} = 0$) et qu'aucun n'était réutilisé ($\gamma_{ij} = 0$).

Le calcul de V_{IMP}^{IMP} découle directement du lemme 2 et les poids sont traités comme des constantes dans l'espérance conditionnelle. Si nous substituons w_i^j à w_i dans l'équation (13), nous obtenons

$$V_{IMP}^{IMP} = 2 \sum_{g=1}^G \left\{ \sum_{i \in A_M^g} w_i^2 \phi_g^2 + \sum_{i, j \in A_M^g} w_i^j \gamma_{ij} \phi_g^2 \right\} + 2 \sum_{g=1}^G \left\{ \sum_{i \in A_M^g} w_i^2 \phi_g^2 + \sum_{i, j \in A_M^g} w_i^j \gamma_{ij} \phi_g^2 \right\}.$$

V_{IMP}^{IMP} ne dépend pas de ce que les donneurs soient intérieurs ou extérieurs au domaine.

Le calcul de V_{MIX}^{MIX} découle aussi des indications de la section 3. Si nous substituons w_i^j à w_i dans l'équation (15), nous obtenons

$$V_{MIX}^{MIX} = \sum_{g=1}^G \left(\sum_{i \in A_M^g} w_i^j d_{ij} - \sum_{j \in A_M^g} w_j^2 \right) \phi_g^2 = \sum_{g=1}^G \left(\sum_{i \in A_M^g} w_i^j d_{ij} - \sum_{j \in A_M^g} w_j^2 \right) \phi_g^2. \quad (20)$$

Il convient de noter que la composante mixte n'est pas nulle pour un total de domaine même en cas d'égalité de toutes les valeurs initiales de pondération. En équilibrage de w (mais sans égalité pour w), la contribution à V_{MIX}^{MIX} est nulle si le donneur est intérieur au domaine et négative s'il y est extérieur. Ainsi, $V_{MIX}^{MIX} = -w^2 \sum_{g=1}^G l_{g^v}^2 \phi_g^2$, étant le nombre de donneurs extérieurs au domaine dans la cellule g . Dans ce cas, si nous ne tenons pas compte de la composante mixte dans l'estimation par domaine, nous nous trouvons à surestimer la variance totale. À poids inégaux, le biais de non-prise en compte de cette composante peut être soit positif soit négatif.

Dans le modèle à moyennes de cellules, la variance totale d'un estimateur (linéaire) de domaine d'imputation est ainsi estimée :

$$V_{TOT}^{TOT} = V_0 + 2 \sum_{g=1}^G \sum_{i < j} w_i^j w_j^j \gamma_{ij} \phi_g^2 + 2 \sum_{g=1}^G \sum_{i \in A_M^g} w_i^j w_j^j d_{ij} \phi_g^2. \quad (21)$$

En guise d'illustration, prenons le cas d'une équilibrage-ratation dans le domaine ($w_{ij}^v = w^v$) et d'une non-réutilisation de (21) est nul et le troisième traduit l'accroissement de la variance d'imputation. Si toutes les valeurs manquantes sont imputées par donneurs intérieurs au domaine, le troisième terme est $2w^v \sum_{g=1}^G m_{g^v}^2 \phi_g^2$, où m_{g^v} est le nombre de

Si la distribution de y varie par domaine (c'est-à-dire que le modèle d'imputation est mal spécifié), le choix de donneurs hors domaine fait que les estimations seront entachées d'un biais. Comme tous les modèles sont plus ou moins mal spécifiés, il serait généralement peu sage de délibérément choisir des donneurs extérieurs au domaine pour minimiser la variance.

5. ÉTUDE DE SIMULATION

Nous avons procédé à une petite étude de simulation afin d'examiner les estimations de variance par l'approche assistée d'un modèle aux fins de l'estimation d'un total général et d'un total de domaine. Nous avons tiré d'une superpopulation infinie un échantillon de 40 grappes comportant précisément 5 unités chacune; y_{mi}^a était la variable d'intérêt pour l'unité i de la grappe a . Nous avons produit les valeurs y par $y_{mi}^a = \tau a^a + e_{mi}^a$ où a^a et e_{mi}^a sont des tirages aléatoires indépendants sur la distribution normale type. Ainsi, les valeurs y sont d'une moyenne nulle, d'une variance $(\tau^2 + 1)$ et d'une corrélation $\rho = \tau^2 / (1 + \tau^2)$ si les unités viennent de la même grappe et $\rho = 0$ dans les autres cas. Nous avons choisi des valeurs de $\tau = 0$ et $\tau = 0,5$, d'où des corrélations respectives de 0 et 0,2. Nous avons retenu la valeur $\rho = 0,2$ pour illustrer l'effet d'une étroite corrélation intracasse. Outre la variable y , nous avons produit une variable indicatrice du domaine v par échantillonnage indépendant pour une probabilité de 0,25 d'appartenance au domaine. Nous avons sélectionné les répondants dans tout l'échantillon avec une probabilité uniforme de réponse de 0,6 et attribué les valeurs manquantes par imputation à donneur intracellulaire et avec remise. Nous avons prélevé 5 000 échantillons de Monte Carlo au total.

Les estimateurs ponctuels en simulation du total général et du total de domaine sont exempts de biais. Le tableau 2 présente les moyennes et les biais des estimateurs de variance par l'approche assistée d'un modèle (V_{TOT}^{TOT}) (les biais relatifs des estimateurs de variance sont très petits, si $\rho = 0,2$, les estimateurs comportent des biais relatifs pour le total général et les totaux de domaine. Par ailleurs, les biais relatifs des estimateurs de variance sont très petits, si $\rho = 0$, les estimateurs sont divisés par $N^2 10^{-4}$). Si $\rho = 0$, les

par rapport au nombre prévu de ces réductions (qui serait le résultat inconditionnel). La stratégie est assez souple pour admettre une diversité de méthodes d'imputation avec ou sans remise et avec ou sans pondération.

4. ESTIMATION PAR DOMAINE

Dans cette section, nous examinerons l'important problème d'estimation par domaine pour les modèles à moyennes de cellules d'une imputation par donneur. Il y a eu peu de recherches qui aient antérieurement été consacrées à ce problème (Lee et coll. 1995). L'estimateur type de réponse complète d'un total de population pour le domaine v est $\hat{\theta}_{n_y} = \sum_{i \in A} w_i y_i$, ce qui peut aussi s'exprimer par $\hat{\theta}_{n_y} = \sum_{i \in A} w_i y_i$, où $w_i = \delta_{vi}$ avec $\delta_{vi} = 1$ si $i \in A_v$ et dans les autres cas, $\delta_{vi} = 0$. L'estimateur imputé par donneur est $\hat{\theta}_{n_y} = \sum_{i \in A} w_i \delta_{vi} y_i = \sum_{i \in A} w_i y_i$. Tout au long, nous posons que δ_{vi} est connu pour tout $i \in A$.

Dans le modèle à moyennes de cellules, nous supposons que tous les éléments d'une cellule sont d'une même distribution. En général, ces éléments feront tantôt partie du domaine, tantôt non. Dans une de ses versions, nous posons l'existence d'un modèle séparé à moyennes de cellules pour le seul domaine et appliquons ensuite un plan d'imputation approprié. La théorie exposée à la section précédente englobe ce cas, et il n'y a donc pas lieu de la reprendre ici. S'il est possible de tenir compte des principaux domaines au stade de l'imputation, il demeure impossible de retenir tous les domaines que l'analyste pourrait vouloir étudier. Nous nous attacherons donc aux domaines intercellulaires, ce qui a d'importantes conséquences pratiques, plus particulièrement dans l'analyse de fichiers de données à grande diffusion.

Examinons maintenant l'estimation des trois composantes de la variance totale V_{TOT}^2 , c'est-à-dire la variance d'un total de domaine d'imputation. Voyons d'abord l'estimation de V_{SAM}^2 . En cas de réponse complète, si nous posons $y_i = 0$ pour les éléments extérieurs au domaine, la variance d'échantillonnage estimée peut s'exprimer par l'équation (7) comme $V_{n_y}^2 = \sum_{i \in A} w_i^2 y_i^2 + 2 \sum_{i < j} w_i w_j y_i y_j$. Comme la composition du domaine est connue pour tous les éléments de l'échantillon, le biais conditionnel de l'estimateur de variance d'imputation V_0 sera ce que nous avons

$$E_g(V_0 - V_{n_y} | \mathbf{A}, \mathbf{A}_R, \mathbf{p}) = 2 \sum_{g=1}^G \sum_{i \in A_R^g} \sum_{j \in A_M^g} w_i w_j \sigma_g^2 + 2 \sum_{g=1}^G \sum_{i < j} w_i w_j \sigma_g^2. \quad (19)$$

Comme nous l'avons aussi dit dans cette section, on peut se servir commodément de V_0 dans le cas de grands échantillons pour estimer V_{SAM}^2 , car on peut utiliser les projections

$$V_{TOT}^2 = V_0^2 + 2 \sum_{g=1}^G \sum_{i \in A_R^g} \sum_{j \in A_M^g} w_i w_j d_{ij} \sigma_g^2 + 2 \sum_{g=1}^G \sum_{i < j} w_i w_j d_{ij} \sigma_g^2. \quad (17)$$

Dans l'examen de cet estimateur, nous citerons quelques exemples simples aux solutions connues. Tous ces exemples visent des échantillons à équipondération où la composante mixte est nulle. Posons d'abord l'existence d'un échantillonnage aléatoire simple avec remise et d'une imputation par donneur avec le modèle général à moyennes de cellules, ainsi que l'absence de réduction de donneurs. Avec le simple estimateur de variance de V_{SAM}^2 , la variance totale estimée est $n^{-1} s_y^2 + 2n^{-1} \hat{\sigma}_2^2 (1 - m^{-1})$, où $s_y^2 = (n-1)^{-1} \sum_{i \in A} (y_i - \bar{y})^2$, \bar{y} est le nombre de répondants et m le nombre de cas manquants. Si nous utilisons $\hat{\sigma}_2^2$ au lieu de s_y^2 (où $\hat{\sigma}_2^2$ est sans biais de modèle, alors que s_y^2 est entaché d'un petit biais d'échantillon), nous pourrions simplifier en $r^{-1} \hat{\sigma}_2^2 [1 + m(r-m)^{-1}]$. En prenant l'espérance de cet estimateur imputé par donneur sans remise que présente l'estimateur (1983, page 25, 2.3.1.7).

Si nous employons un modèle multiple au lieu d'un modèle général à moyennes de cellules, la variance totale estimée est $n^{-1} s_y^2 + 2n^{-1} \sum_{g=1}^G \hat{\sigma}_2^2 (n_g^g - r_g^g)$, ce qui correspond au résultat présenté par Tolleson et Fuller (1992). Dans le même exemple d'un échantillonnage aléatoire simple, permettons maintenant une réduction de cellules. La encore, si nous prenons $\hat{\sigma}_2^2$ au lieu de s_y^2 , la variance totale estimée est approximativement

$$n^{-1} \hat{\sigma}_2^2 (n + m + \sum_{i < j} w_i w_j) + \sum_{i \in A_M^g} w_i^2 \sigma_g^2. \quad (18)$$

Pour m fixe, la variance est minimisée dans l'équation (18) lorsqu'aucun donneur n'est utilisé plus souvent que tout autre dans la mesure du possible (on minimise de ce fait le biais de l'estimateur). Ainsi, un mode d'imputation où un donneur est utilisé au plus une fois de plus que tout autre se trouve à minimiser la variance totale. Si nous sélectionnons les donneurs par échantillonnage aléatoire simple avec remise, $E_g[V_{TOT}^2 | \mathbf{y}] = r^{-1}$ et l'espérance de la variance de l'estimateur imputé par donneur avec remise que présente Kalton (1983, page 26, 2.3.1.9).

Les exemples indiquent que ce traitement produit des estimations raisonnables de la variance totale dans des cas simples et fait bien voir le caractère conditionnel des estimations de variance. Ainsi, (18) est conditionnel par rapport au nombre effectif de réductions de donneurs, et non pas

Démonstration. Nous commençons par écrire $(\theta_I - \theta_n) (\theta_n - \theta_n) = \theta_n (\theta_I - \theta_n) - \theta_n (\theta_I - \theta_n)$. Soit θ_n le total de la population finie sous la forme $\sum_{i \in U-A} Y_i + \sum_{i \in A} X_i$. Dans cette expression, le second élément peut ainsi être développé :

$$\theta_n (\theta_I - \theta_n) = \left(\sum_{i \in U-A} Y_i + \sum_{i \in A} X_i + \sum_{j \in A} Y_j \right) \left[\sum_{j \in A} w_j (Y_j^* - Y_j) \right].$$

Si nous prenons l'espérance conditionnelle de ce produit, les seules contributions non nulles se présentent lorsque l'unité i de A_R est le donneur de Y_j^* ou que l'unité i de A_M dans le premier ensemble entre parenthèses est l'unité j de l'autre ensemble. Dans le premier cas, $E_{\pi} (Y_i Y_j^* - Y_j) = d_{ij} \sigma_2^2$ pour $i \in A_R, j \in A_M$. Dans le second cas, si l'unité non répondante i de A_M est la même que l'unité j du second terme, $i = j$, $E_{\pi} (Y_i (Y_j^* - Y_j)) = -\sigma_2^2$, cette espérance étant nulle dans les autres cas. Ainsi,

$$E_{\pi} (\theta_n (\theta_I - \theta_n) | \mathbf{A}, \mathbf{A}_R, \mathbf{d}) = \sum_{j \in A_M} \sum_{i \in A_R} w_i w_j \sigma_2^2 - \sum_{j \in A_M} w_j \sigma_2^2 = 0.$$

Le premier terme peut ainsi s'exprimer :

$$\theta_n (\theta_I - \theta_n) = \left(\sum_{i \in A_R} w_i Y_i + \sum_{i \in A_M} w_i Y_i \right) \left(\sum_{j \in A_M} w_j (Y_j^* - Y_j) \right).$$

Par les résultats pour $E_{\pi} (Y_i (Y_j^* - Y_j))$ plus haut,

$$V^{MIX} = \sum_{g=1}^G \sum_{i \in A_R, j \in A_M} w_i w_j d_{ij} - \sum_{g=1}^G \sum_{i \in A_M} w_i^2 \sigma_g^2 \quad (16)$$

L'estimateur de V^{MIX} est nul lorsque les poids sont constants ou, plus généralement, que les poids des donneurs correspondent aux poids des cas manquants auxquels leurs valeurs s'appliquent. Dans la plupart des simulations décrites dans les études spécialisées (voir, par exemple, Samdal 1992, et Lee, Rancourt et Samdal 1995), on se sert d'échantillons aléatoires simples de sorte que les estimations du terme mixte de ces simulations soient approximativement égales à zéro.

Pour illustrer l'effet d'inégalité de pondération, considérons un échantillon aléatoire simple stratifié tiré de deux strates avec remise. Posons aussi que le taux d'échantillonnage est k fois plus élevé dans la strate 2 que dans la strate 1. Supposons enfin que le modèle d'imputation est le modèle général à moyennes de cellules et que la technique d'imputation par donneur sélectionne les donneurs par

échantillonnage aléatoire simple sans remise. Dans ce cas simple, V^{MIX} peut se calculer algébriquement. Le tableau 1 indique la contribution en pourcentage du terme mixte à la variance totale ($100 \cdot 2V^{MIX} / V^{TOT}$) pour diverses combinaisons de taux de réponses des strates. Il illustre le fait que, en cas d'inégalité de pondération, la contribution de ce terme peut être importante ainsi que positive ou négative. Le terme mixte peut aussi importer dans une estimation par domaine. Nous en parlerons à la prochaine section.

Tableau 1
Contribution en pourcentage du terme mixte à la variance totale V^{TOT}

Taux de réponse	Taux de suréchantillonnage de la strate 2		Taux de réponse		Taux de suréchantillonnage de la strate 2	
	Strate 1	Strate 2	Strate 1	Strate 2	Strate 1	Strate 2
100%	80%	4,3	5,0	13,7	18,3	13,7
100	60	8,7	10,8	17,7	29,7	17,7
100	40	13,7	18,3	28,8	44,5	28,8
60	100	-15,4	-34,1	-27,1	-37,6	-27,1
60	80	-10,4	-27,1	-19,0	-29,3	-19,0
60	60	-5,2	-19,0	-8,8	-18,2	-8,8
60	40	1,0	-8,8	6,5	0,0	6,5
60	20	9,4	6,5			

ne soit pas extrêmement grand dans chaque cellule. Pour les deux autres composantes, les seules quantités inconnues à estimer à partir de l'échantillon sont les variances de cellules σ_g^2 . Il est possible d'estimer ces paramètres par des observations avec ou sans pondération, les valeurs de pondération appliquées étant les poids de sélection. Fuller (2002) recommande des observations pon-détées pour des estimations plus robustes. Nous calculons les estimateurs sans biais de la variance conditionnelle d'imputation et de la composante mixte par substitution des estimations sans biais des variances de cellules σ_g^2 . L'addition de V_0^{IMP} et $2V^{MIX}$ donne alors l'estimateur de la variance totale

du biais du simple estimateur de variance. À noter que, toutes les fois que des répondants donnent des valeurs à plus d'un non-répondant, le dernier terme de l'équation (12) est positif et que, dans les autres cas, il est nul.

Deux exemples simples illustreront l'application de ces résultats. Considérant d'abord l'estimation d'une moyenne de population à partir d'un échantillon aléatoire simple avec remise. Dans ce cas, $\Omega_{ij}^g = n^{-2}$ et $\Omega_{ij}^g = -n^{-2}(n-1)^{-1}$ pour $i \neq j$. Posons que le modèle à moyennes de cellules est valable dans une imputation par donneur et qu'aucun donneur n'est utilisé plus d'une fois. Selon le lemme 1, le biais de V_0^g est $-2n^{-2}(n-1)^{-1} \sum_{i \in A_M^g} m_i^g \sigma_{ij}^2$ où $m_i^g = \sum_{j \in A_M^g} d_{ij}^g$ est le nombre de valeurs imputées dans la cellule g , auquel cas le biais du simple estimateur de variance est $O_p(n^{-2})$ et donc négligeable pour un n élevé. Supposons maintenant que, dans chaque cellule, toute valeur manquante est imputée par le même donneur. Dans ce cas, comme $\sum_{i < j \in A_M^g} y_{ij}^g = m_i^g(m_j^g - 1)/2$, le biais de V_0^g est $-(n^{-2}(n-1)^{-1} \sum_{i \in A_M^g} (m_i^g)^2 + m_i^g) \sigma_{ij}^2$ d'où $O_p(n^{-1})$ et une valeur du même ordre que la variance d'échantillonnage.

Comme autre exemple, prenons un échantillon simple à deux degrés de taille $n = ab$ où a grappes sont tirées d'une population de A grappes de taille égale par échantillonnage aléatoire simple et où b éléments sur B sont également sélectionnés par échantillonnage aléatoire simple dans chaque grappe échantillonnée. Soit y_{ai}^g la valeur de y pour l'unité sélectionnée i de la grappe a . Posons que, au premier degré, le taux d'échantillonnage est suffisamment petit pour qu'on n'en tienne pas compte. L'estimation de la variance de la moyenne d'échantillon est de la forme donnée par l'équation (7), où $\Omega_{a,b,f}^g = a^{-2}b^{-2} = n^{-2}$ pour $a = b$ et où $\Omega_{a,b,f}^g = -n^{-2}(a-1)^{-1}(b-1)^{-1}$ pour $a \neq b$. Nous pouvons maintenant insérer ces valeurs dans l'équation (9) pour estimer le biais. Supposons, par exemple, que toutes les valeurs manquantes sont imputées par donneur de la même grappe (nos cellules sont ces grappes) et qu'aucun donneur n'est utilisé plus d'une fois. Dans ce cas, le biais du simple estimateur de variance est $2n^{-2} \sum_a m_a^g \sigma_{a,a}^2$ où m_a^g est le nombre de non-répondants de la grappe a . Posons qu'une imputation générale par donneur par l'approche assistée d'un modèle à moyennes de cellules est effectuée et qu'aucun donneur n'est réutilisé, mais que les donneurs sont toujours tirés de grappes différentes de celles des valeurs manquantes correspondantes. Dans ce cas, le biais du simple estimateur de variance est $-2n^{-2}(a-1)^{-1} \sum_a m_a^g \sigma_{a,a}^2$. Cet exemple de traitement à deux degrés indique que cet estimateur peut être biaisé dans l'un ou l'autre sens, que, dans les deux cas considérés, il est d'un ordre inférieur à celui de la variance et que, si a est élevé, il sera négligeable. La deuxième composante de la variance totale est la variance d'imputation V_{IMP}^g . Le lemme 2 présente un

estimateur sans biais de cette composante pour l'imputation par donneur.

Lemme 2. Pour des hypothèses semblables à celles du lemme 1, un estimateur sans biais de V_{IMP}^g est

$$V_{IMP}^g = 2 \sum_{i < j \in A_M^g} \sum_{i \in A_M^g} w_i^g \hat{\sigma}_{ij}^g + \sum_{i \in A_M^g} w_i^g y_{ij}^g \hat{\sigma}_{ij}^g \quad (13)$$

où $\hat{\sigma}_{ij}^g$ est un estimateur sans biais de σ_{ij}^g .

Démonstration. Comme la variance due à l'imputation est la différence quadratique entre les estimations d'imputation et de réponse complète, nous commencerons par écrire

$$\begin{aligned} & (\theta_j - \theta_n)^2 = \left[\sum_{i \in A_M^g} w_i^g (\tilde{y}_i^* - y_i^*) \right]^2 \\ & = \sum_{i \in A_M^g} w_i^g (y_i^* - y_i)^2 + 2 \sum_{i < j \in A_M^g} w_i^g w_j^g (y_i^* - y_i)(y_j^* - y_j) \\ & \quad + 2 \sum_{i < j \in A_M^g} w_i^g w_j^g (y_i^* - y_i)(y_j^* - y_j) \end{aligned}$$

Il convient de noter que $E_g(y_i^* - y_i)^2 = 2\sigma_{ij}^g$ pour i de la cellule g et que, comme $E_g[(y_i^* - y_i)(y_j^* - y_j)] = E_g(y_i^* y_j^* - y_i y_j^* - y_i^* y_j + y_i y_j) = y_{ij}^g \sigma_{ij}^g$, il s'ensuit que

$$V_{IMP}^g = 2 \sum_{i < j \in A_M^g} w_i^g w_j^g \sigma_{ij}^g + \sum_{i < j \in A_M^g} w_i^g w_j^g y_{ij}^g \sigma_{ij}^g \quad (14)$$

En substituant $\hat{\sigma}_{ij}^g$, estimateur sans biais de σ_{ij}^g , nous établissons ce lemme.

L'équation (14) indique que la variance d'imputation reçoit une contribution positive de chaque valeur imputée et de la réutilisation de donneurs. Supposons, par exemple, que les valeurs de pondération sont égales pour tous les cas échantillonnés. La contribution de la cellule g à cette variance est alors proportionnelle à la somme des nombres respectifs de cas manquants dans cette cellule et de paires de non-répondants qui reçoivent leurs valeurs des mêmes donneurs. En restreignant le nombre de réutilisations de donneurs, on peut réduire la variance d'imputation.

La troisième composante de la variance totale est V_{MIX}^g . que, dans des recherches antérieures, on a souvent considérée comme petite, voire négligeable (voir, par exemple, Särndal 1992, et Deville et Särndal 1994). Le lemme 3 donne un estimateur sans biais de V_{MIX}^g .

Lemme 3. Pour des hypothèses semblables à celles du lemme 1, un estimateur sans biais de V_{MIX}^g est

$$\hat{V}_{MIX}^g = \sum_{i < j \in A_M^g} \sum_{i \in A_M^g} w_i^g w_j^g d_{ij}^g - \sum_{i \in A_M^g} w_i^g \hat{\sigma}_{ij}^g \quad (15)$$

[illegible]

pour des coefficients connus $\Omega^{\tilde{H}}$. Cette formulation comprend l'estimateur d'Horvitz-Thompson où les $\Omega^{\tilde{H}}$ sont

déterminées par les probabilités simples et composées de sélection. Elle comprend aussi l'estimateur de variance linéarisé pour l'estimateur généralisé de régression (EGR).

linéarisé peut s'écrire par substitution de $g_{is}^0 e_i^j$ à y_i^j dans l'estimateur de variance pour l'estimateur d'Horvitz-

dépendance de l'échantillon et $e_i = y_i - \hat{x}_i' \hat{\beta}$, où \hat{x}_i est le vecteur de variables auxiliaires et $\hat{\beta}$, le vecteur de

est linéaire dans y . Ainsi, l'estimateur linéarisé de variance

que, dans ce cas, les Ω_{ij}^{η} peuvent être en dépendance de

Devil et Sarnad (1992) indiquent que tout estimateur d'étalonnage comporte la même variance asymptotique que

asymptotique pour les estimateurs d'étalement en général ont la forme quadratique requise.

Le simple examen de variance pour les valeurs imputées comme valeurs d'observation et peut ainsi s'écrire :

Le lemme 1 donne le biais du simple estimateur de variance en tant qu'estimateur de V . Comme nous l'avons

signalé, il doit, comme estimateur de V_{SAM} , convenir le

« naïf » est de vouloir mettre à profit l'existence sur le marché de progiciels qui estiment la variance d'échantillonnage dans des plans de sondage complexes.

Lemme 1. Dans le modèle à moyennes de cellules où les mécanismes d'échantillonnage, de réponse et d'imputation sont « non confondus » et où on pose que θ est un

estimation impuise linéaire sans biais par donneur comme dans (6) et que \hat{V}_n est un estimateur sans biais de variance échantillonale complet comme dans (7), le biais du simple \hat{V}_0 est, comme estimateur de V :

$$(01) \quad p^{12}p^3 \sum_{\lambda}^{\mu_{V \ni \lambda}} = h_{\lambda}$$

Démonstration. Nous commençons par noter que la différence entre V_0^i et V^n peut ainsi s'écrire :

$$\left(\frac{1}{z}K - \frac{1}{z_{\sim}}K\right)^n \mathfrak{U} \sum_{\forall \exists !} = {}^u A - {}^0 A$$

$$\left({}^f\chi^1_\sim - {}^f\chi^1_\sim \right) \mathfrak{U} \mathfrak{Z} \mathfrak{Z} \mathfrak{z} +$$

$$\left(\frac{1}{z} \kappa - \frac{1}{z^*} \kappa^*\right) \mathfrak{U}^{\mathfrak{A} \in \mathfrak{I}} \mathfrak{Z} =$$

$$\left(\int_{\mathcal{K}} \chi - \int_{\mathcal{K}^*} \chi \right) \int_{\mathcal{V}} \chi \geq 0 \quad \forall \chi \in \mathcal{V}, \int_{\mathcal{V}} \chi \geq 0$$

$$(II) \quad \cdot \left(\sum_{\substack{f > i \\ f \in M}} \chi^{\dagger} \chi - \sum_{*} \chi^{\dagger} \chi \right) \bar{v}_2 \bar{v}_3 v_+$$

Dans le modèle à moyennes de cellules, l'estimation conditionnelle du premier terme de (11) est nulle.

$E_{\xi}^{\varepsilon}[y_i(y_i^* - y_j)] = 0$ sauf si le répondant i est le donneur du non-répondant j . Elle est donc nulle lorsque les unités i et j

g. Elle peut être représentée par $E_{ij}^g[y_i^g(y_j^g - y_j^g)] = d_{ij}^g \sigma_{ij}^g$ de zéro uniquement pour des unités i et j de la même cellule

qui se produira seulement si ces unités appartiennent à la

$E_{\hat{\zeta}}^{\zeta}(y_i^* y_j^* - y_i y_j) = \gamma_{ij}^{\zeta}$ pour $i \neq j$. Si nous appliquons ces résultats à l'équation (11), nous obtenons

$$\frac{\partial}{\partial \mathbf{p}} \sum_{\mathbf{v} \in \mathcal{V}} \sum_{\mathbf{v}' \in \mathcal{V}} \sum_{\mathbf{v}'' \in \mathcal{V}} \mathcal{L} = \left(\mathbf{p}, \mathbf{A}^R, \mathbf{p} \mid \mathbf{A}, \mathbf{A}^R, \mathbf{p} \right) \left(\mathcal{L}^0 - \mathcal{L}^n \right) \left(\mathcal{L}^{\mathbf{v}''} \right)$$

$$(12) \quad \sum_{\alpha} \sum_{l=1}^{\infty} \sum_{j \geq l} \tau_j \sigma_2^{(j)} \cdot \omega_{\lambda}^{(j)}.$$

La démonstration est doublee lorsque nous disons que, comme V^n est sans biais selon le plan de sondage, il est aussi pour $V^{S\Delta M}$. La substitution d'un estimateur sans biais de σ^2 donne un estimateur sans biais de σ^2 .

2. IMPUTATION PAR DONNEUR

Nous considérerons un modèle simple auquel une imputation par donneur est applicable. Posons que la population finie (U) est formée de G classes ou cellules. Dans la cellule g ($g = 1, \dots, G$), les éléments de U sont des réalisations de variables indépendantes et identiquement distribuées d'une moyenne μ_g et d'une variance σ_g^2 . Ce modèle à moyennes de cellules peut ainsi s'écrire :

$$(4) \quad X_{it} \mid (\mu_g, \sigma_g^2), i \in U_g.$$

où i_t est l'abréviation de l'expression « variables indépendantes et identiquement distribuées ».

Un estimateur linéaire de θ_n en cas de réponse complète à une enquête complexe peut se présenter sous la forme suivante :

$$(5) \quad \theta_n = \sum_{i \in A} w_i y_i,$$

où w_i est la valeur de pondération qui tient compte des probabilités inégales de sélection et la stratégie d'estimation. Là où le modèle à moyennes de cellules est valable, un estimateur plus efficace de θ_n prend les moyennes de groupes sans pondération, $\theta_n^* = \sum_{i \in A} \sum_{j \in A_g} w_i y_j$, où $y_j^* = y_j^g/n_g$. Toutefois, l'approche assistée d'un modèle ne s'appuie pas complètement sur le modèle employé : on s'en tient au traitement de pur plan de sondage dans la mesure du possible et n'applique le modèle qu'aux données manquantes. Les valeurs de pondération de (5) peuvent être l'inverse des poids des probabilités de sélection ou de correction d'échantillonnage, ainsi que nous allons le décrire.

La valeur d'imputation par donneur de y_j est $y_j^* = \sum_{i \in A_g} d_{ij} y_i$ et l'estimateur imputé est

$$(6) \quad \theta_I = \sum_{i \in A} w_i y_i + \sum_{j \in A_g} w_j \sum_{i \in A_g} d_{ij} y_i,$$

où $y_i = y_i$ pour $i \in A_g$ et $y_i = y_i^*$ pour $i \in A_M$. Nous posons tout au long que les valeurs d'imputation viennent d'unités répondantes de la même cellule et que chaque cellule contient au moins une de ces unités.

Dans la formation de cette imputation, nous ne spé-
cifierons pas le mode de sélection des donneurs. Sont donc visées les imputations par donneur sans pondération où les probabilités de sélection d'unités sont égales dans chaque cellule, ainsi que les imputations avec pondération. On se sert habituellement de ces imputations en pondération lorsque les hypothèses posées portent seulement sur la distribution de la réponse. La formule (6) vise aussi les méthodes d'imputation avec ou sans remise. Ainsi, une imputation répandue est celle où un répondant est choisi au hasard comme donneur dans une cellule, mais n'est plus

utilisé par la suite tant que tous les autres répondants de la cellule ne l'ont pas été.

Nous n'en tirons pas expressément ici, mais les modes d'imputation par plus proche voisin où on se sert de variables continues pour établir un petit ensemble de répondants assimilables et ensuite en choisir un au hasard pour imputation répondent tout autant à nos critères. Ajoutons que les chercheurs emploient souvent des méthodes par donneur même lorsqu'ils disposent de variables continues. Little (1986) examine les stratégies de formation de cellules d'imputation à l'aide de variables prévisionnelles de la variable y et fait observer que l'imputation intracellule et l'imputation par régression devraient produire des résultats analogues dans bien des cas. Cochran (1968) et Aigner, Goldberger et Kalton (1975) démontrent pour leur part qu'un nombre relativement modeste de cellules bien construites à partir d'une variable continue peuvent appréhender une grande partie de la puissance pré-

visionnelle de la variable.

Dans le modèle à moyennes de cellules, le biais conditionnel de l'estimateur imputé est

$$E_{\xi}(\theta_I - \theta_n \mid \mathbf{A}, \mathbf{A}_g, \mathbf{d}) = E_{\xi} \left[\sum_{j \in A_g} w_j (y_j^* - y_j) \mid \mathbf{A}, \mathbf{A}_g, \mathbf{d} \right] = 0,$$

puisque $E_{\xi}(y_j^*) = E_{\xi}(\sum_{i \in A_g} d_{ij} y_i) = \sum_{i \in A_g} d_{ij} E_{\xi}(y_i) = \sum_{i \in A_g} d_{ij} \mu_g = \mu_g$ pour j dans la cellule g . Cette espérance est en coïncidence par les indices des unités échantillonnées et répondantes et des donneurs. Toutefois, comme l'estimateur est conditionnellement sans biais pour tout échantillon, il l'est aussi inconditionnellement. Kim et Fuller (1999) usent aussi de cet argument du conditionnement. Nous présenterons à la prochaine section les estimateurs des diverses composantes de la variance de l'estimateur imputé par donneur.

3. ESTIMATION DES COMPOSANTES DE LA

VARIANCE TOTALE

Dans cette section, nous livrons les principaux résultats en ce qui concerne les estimateurs des trois composantes de la variance totale d'un estimateur linéaire d'imputation par donneur. Tout au long, nous poserons que les mécanismes d'échantillonnage, de réponse et d'imputation ont un caractère « non confondu » et qu'il existe un estimateur linéaire d'échantillon complet de la forme (5). Nos résultats exigent que le modèle à moyennes de cellules soit valable et qu'il y ait au moins un répondant dans chaque cellule d'imputation. Nous débuterons par la variance d'échan-

Ainsi, la validité du modèle est un aspect bien plus primordial d'une estimation de variance due à l'imputation par l'approche assistée d'un modèle que d'un échantillonnage assisté d'un modèle. Särndal (1992) soutient que, si nous sommes disposés à accepter la validité du modèle dans une estimation ponctuelle avec données d'imputation, nous devrions l'être autant dans une estimation de variance. Nous obtenons des estimateurs de variance en condition-répondantes et d'imputations. Nous élaborons des estimateurs de $V_{S\Delta M}^{\varepsilon}[(\theta_I - \theta_N)^2 | \mathbf{A}, \mathbf{A}_R, \mathbf{d}]$, et $V_{MIX}^{\varepsilon}[(\theta_I - \theta_N)^2 | \mathbf{A}, \mathbf{A}_R, \mathbf{d}]$.

où \mathbf{A} et \mathbf{A}_R désignent des matrices d'indices pour les unités échantillonnées et répondantes respectivement et où \mathbf{d} est l'ensemble d'indices des imputations. Le conditionnement est par les indices, et non par les valeurs des unités. \mathbf{d} est une matrice $r \times (n - r)$ avec les répondants en ligne et les non-répondants en colonne. Il sera uniquement question ici de méthodes simples d'imputation où tous les d_j sauf un sont égaux à 0 dans chaque colonne. Cette exception se présente à la ligne du répondant donneur où $d_j = 1$.

Par rapport aux espérances conditionnelles de V_{IMP}^{ε} et V_{MIX}^{ε} , les estimateurs tiennent compte du nombre de fois que des unités répondantes servent de donneurs dans l'application. Il ne s'agit donc pas de prendre l'espérance sur l'ensemble des résultats possibles d'imputation. Comme nous le ferons valoir, ce sont les variances qu'il convient d'estimer dans une application. Si les estimateurs de variance sont conditionnellement sans biais, ils le seront aussi inconditionnellement, bien sûr.

Un cadre conditionnel est utile à deux égards. D'abord, si un estimateur est conditionnellement sans biais et converge (on suppose que θ_I l'est pour θ_N), la variance conditionnelle est généralement un estimateur plus approprié que la variance inconditionnelle pour les inférences à tirer de l'échantillon réalisé (Holt et Smith 1979; Rao 1999; Kalton 2002). Ainsi, un estimateur de variance en conditionnant par la fréquence d'utilisation de chaque répondant donneur est à préférer à un estimateur qui prend la moyenne sur tous les choix possibles de donneurs. En second lieu, les résultats s'appliquent à tous les mécanismes d'échantillonnage, de réponse et d'imputation à caractère non confondu qui produisent le même ensemble d'unités échantillonnées et répondantes et d'imputations. Ainsi, les résultats que nous présentons pour l'imputation à caractère non confondu qui attribue des valeurs d'observation là où il y a des valeurs manquantes et dans lequel $E_{\varepsilon}^{\varepsilon}(\theta_I) = E_{\varepsilon}^{\varepsilon}(\theta_N)$.

Par la décomposition de l'équation (1), Särndal (1992) exprime ainsi la variance totale pour l'estimateur imputé :

$$V^{\text{TOT}} = E_{\varepsilon}^{\varepsilon} E_p E_R E_I (\theta_I - \theta_N)^2 = V^{\text{SAM}} + V^{\text{IMP}} + 2V^{\text{MIX}} \quad (3)$$

où $V^{\text{SAM}} = E_{\varepsilon}^{\varepsilon} E_p (\theta_N - \theta_N)^2$ est la variance d'échantillonnage, $V^{\text{IMP}} = E_{\varepsilon}^{\varepsilon} E_p E_R E_I (\theta_I - \theta_N)^2$ la variance d'imputation et $V^{\text{MIX}} = E_{\varepsilon}^{\varepsilon} E_p E_R E_I ((\theta_I - \theta_N) - (\theta_N - \theta_N))$ une composante mixte. Dans cette formulation, la variance totale et ses composantes seraient plutôt des variances de prévision, puisqu'elles tiennent compte de l'espérance supplémentaire du modèle de superpopulation.

Il convient de distinguer l'estimation de variance due à l'imputation par l'approche assistée d'un modèle que nous décrivons de l'échantillonnage assisté d'un modèle (Särndal, Swensson et Wretiman 1992). Dans un tel échantillonnage, des modèles servent au choix de plans de sondage et d'estimateurs efficaces, mais la validité des inférences statistiques ne dépend pas de celle des modèles. En revanche, si des données manquent, il est essentiel de faire dépendre les inférences des modèles tant pour les estimateurs ponctuels que pour les estimateurs de variance. Dans le cadre général d'inférence que nous exposons, nous employons les hypothèses de modèle d'imputation (celles du modèle de superpopulation et de l'état « non confondu » des mécanismes) dans la seule mesure nécessaire à la prise en compte de données d'imputation. Et les estimateurs ponctuels et les estimateurs de variance sont des estimateurs de pur plan de sondage, c'est-à-dire sans données manquantes. Qu'ils soient approximatifs ou non biais pour V^{SAM} dépend de la validité du modèle d'imputation. Il faut aussi dire que les estimateurs de V^{IMP} et V^{MIX} dépendent entièrement de ce modèle.

Les mécanismes à caractère « non confondu » donnent Rubin (1987, page 36-39) examen plus en détail probabiliste de réponse dans chaque cellule d'imputation par valeur de la variable à imputation est indépendante de la ou cellules d'imputation. Ainsi, nous supposons que la exemple, variables de stratification pour l'échantillonnage analysée en conditionnant par les variables auxiliaires (par question sont indépendants de la distribution de la valeur Y cet état non confondu implique que les mécanismes en manière suivante : $V^{\text{TOT}} = E_p E_R E_I E_{\varepsilon}^{\varepsilon} (\theta_I - \theta_N)^2$. En gros, modèle. Ainsi, la variance totale peut se récrire de la manière à pouvoir prendre d'abord l'espérance relative au mécanismes permet de changer l'ordre des espérances de modèle. Pour l'essentiel, l'état « non confondu » des nous avons énumérés à propos de l'approche assistée d'un Särndal (1992) et tous les autres auteurs spécialisés que « non confondu », ainsi que le décrivent Rubin (1987), mécanismes de réponse et d'imputation sont à caractère d'imputation. Nous posons que le plan de sondage et les modèle et des mécanismes d'échantillonnage, de réponse et où $E_{\varepsilon}^{\varepsilon}$, E_p , E_R , et E_I sont les espérances respectives de ce

Estimation de variance pour l'imputation hot deck à l'aide d'un modèle

J. MICHAEL BRICK, GRAHAM KALTON et JAE KWANG KIM¹

RÉSUMÉ

Dans une imputation où on attribue des valeurs à des réponses manquantes d'enquête par sondage, de simples méthodes d'estimation de variance des estimations d'enquête où on traite les valeurs d'imputation comme s'il s'agissait de valeurs observées donnent des estimations entachées d'un biais. Nous traiterons de ce problème dans le cas d'un estimateur linéaire où les valeurs manquantes sont attribuées par simple imputation par donneur (c'est une forme d'imputation qui est répandue dans la pratique). Nous proposerons des estimateurs de variance pour un estimateur linéaire à imputation par donneur (imputation « hot-deck ») en procédant à une décomposition de la variance totale proposée par Särndal (1992). Nous concevons un traitement conditionnel d'estimation de variance qui est applicable à une imputation par donneur avec et sans pondération. Nous examinerons enfin l'estimation de variance pour un estimateur par domaine.

MOTS CLÉS : Données manquantes; approche assistée d'un modèle; estimation de variance conditionnelle.

1. INTRODUCTION

Il se pose dans la pratique un grand problème d'estimation de la variance d'estimations tirées d'un ensemble de données où une partie des réponses manquent et où les valeurs manquantes sont attribuées par imputation. On a abordé ce problème de bien des façons (voir, par exemple, Rubin, 1987, et Rao et Shao, 1992). Notre propre traitement est fondé sur l'approche assistée d'un modèle proposée par Särndal (1992). Dans son application initiale, Särndal a eu recours à l'approche assistée d'un modèle avec un échantillon aléatoire simple où l'imputation était déterministe et se faisait par rapports. Par la suite, il s'est intéressé à d'autres modes d'imputation et plans de sondage (voir, par exemple Deville et Särndal 1994; Rancourt et Lee et Rancourt, Deville et Särndal 1996). Dans cet article, nous étendons l'approche assistée d'un modèle à des formes générales d'estimateurs linéaires où l'imputation se fait par donneur de la même cellule d'imputation. Une imputation de ce genre où on remplace une valeur manquante par la valeur observée d'une unité répondante de la même cellule est une des méthodes les plus répandues dans le cas des enquêtes par sondage auprès des ménages (Brick et Kalton 1996). C'est par un cadre conditionnel que nous établirons, pour des estimateurs d'imputation par donneur, un estimateur de variance qui vaut pour les plans de sondage généraux et une diversité de stratégies d'estimation.

Dans l'approche assistée d'un modèle, nous exprimons ainsi la différence entre un estimateur imputé (terme que nous employons pour désigner un estimateur reposant en partie sur des valeurs imputées) θ_I et le paramètre correspondant de population finie θ_N :

$$V^{\text{TOT}} = E_{\xi} E_{\theta}^p E_{\theta} E_{\theta_I} (\theta_I - \theta_N)^2, \quad (2)$$

d'un estimateur imputé est

Dans le modèle de superpopulation, la variance totale traitement dans l'estimation de variance.

L'approche assistée d'un modèle pose des hypothèses au sujet de la distribution de la variable d'intérêt dans la population. Ces hypothèses forment un modèle de superpopulation désigné par ξ . En général, l'imputation vise à créer un ensemble de données polyvalent se prêtant à des analyses valides des plus différentes avec peut-être des associations entre une variable à imputation et toute autre variable de l'ensemble de données. Comme il nous faut un modèle de superpopulation pour l'imputation de réponses manquantes à des questions où ces associations seront conservées, il est naturel de recourir aussi à ce traitement dans l'estimation de variance.

quantas.

L'approche assistée d'un modèle pose des hypothèses au sujet de la distribution de la variable d'intérêt dans la population. Ces hypothèses forment un modèle de superpopulation désigné par ξ . En général, l'imputation vise à créer un ensemble de données polyvalent se prêtant à des analyses valides des plus différentes avec peut-être des associations entre une variable à imputation et toute autre variable de l'ensemble de données. Comme il nous faut un modèle de superpopulation pour l'imputation de réponses manquantes à des questions où ces associations seront conservées, il est naturel de recourir aussi à ce traitement dans l'estimation de variance.

$$\theta_I - \theta_N = \left(\theta_I^n - \theta_N^n \right) + \left(\theta_I^n - \theta_I^n \right), \quad (1)$$

où θ_I^n est l'estimateur courant et approximativement sans biais de plan de sondage de θ_N^n en cas de réponse complète. Le premier terme du côté droit de (1) s'appelle l'erreur d'échantillonnage et dépend uniquement de la distribution d'échantillonnage de l'estimateur selon le plan de sondage ayant servi au tirage de l'échantillon complet, ce que nous désignons par p . Le second terme est l'erreur d'imputation qui dépend de la distribution d'échantillonnage, du mécanisme de réponse (R) qui dégage le sous-ensemble de répondants de tout l'échantillon et du mécanisme d'imputation (I) servant à l'attribution des valeurs manquantes. Il ne sera question dans notre exposé que d'estimateurs θ_I pour une seule variable pouvant avoir des données manquantes.

¹ J. Michael Brick et Graham Kalton, Westat, 1650, boulevard Research, Rockville, Maryland 20850, États-Unis, Courriel : mikebrick@westat.com; Jae Kwang Kim, Department of Applied Statistics Yonsei University, Seoul 120-749, Corée.

$$|\phi_i - \bar{\phi}_g| = |\phi^*(c) \parallel X_i - X_{0g}| \leq C \frac{G_v}{X^{(N)} - X^{(1)}} \quad (8)$$

pour une constante $C \in (0, \infty)$ et, en vertu de (A5) et (A6),

$$\left| \text{Biais} \left(\tilde{\tau}_{wc}^g \right) \right| \leq C \gamma_6^{-1} \gamma_5 \frac{G_v}{X^{(N)} - X^{(1)}}.$$

Observons maintenant que, puisque

$$|\tilde{Y}_{ig}^*| \leq \frac{1}{\phi_i} \frac{\phi_g}{\phi_i} |X_i| + \frac{\pi_i}{1} \frac{\phi_g}{\phi_i - \bar{\phi}_g} \left| \tilde{Y}_g^* \right|,$$

alors, en vertu de (A1), (A6) et (8),

$$\tilde{Y}_{ig}^* = O \left(\frac{n_v}{N_v} \right) + O \left(\frac{n_v G_v}{N_v} \right), \quad \forall U_g, \forall g = 1, 2, \dots, G_v,$$

ce qui implique que

$$\tilde{Y}_{ig}^* \tilde{Y}_{jg}^* = O \left(\frac{n_v^2}{N_v^2} \right) + O \left(\frac{n_v^2 G_v^2}{N_v^2} \right), \quad \forall U_g, \forall g = 1, 2, \dots, G_v.$$

Puisque, selon (A7), $N_g^*/N_v = O(1/G_v)$, selon (A2) et (A3), $\sum_{i=1}^{U_g} \Delta_{ij}^* = O(n_v/G_v)$ et, pour $g \neq g'$, $\sum_{i=1}^{U_{g'}} \Delta_{ij}^* = O(n_v/G_v^2)$, alors, le premier terme de $\text{Var}(\tilde{\tau}_{wc}^g/N_v)$ est borné par

$$O \left(\frac{1}{1} \right) + O \left(\frac{n_v G_v}{1} \right).$$

Puisque le deuxième terme de $\text{Var}(\tilde{\tau}_{wc}^g/N_v)$ est borné par $O(1/n_v)$, la conclusion s'ensuit.

BIBLIOGRAPHIE

- CASSELL, C.-M., SÄRNÄL, C.-E. et WRETMAN J.H. (1983). Some uses of statistical models in connection with the *Theory and bibliographies*, (Eds. W.G. Madow, I. Olkin et D. B. Rubin). Academic Press, New York: London. 3, 143–160.
- COCHRAN, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 24, 295–313.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3^{ème} éd.). New York: John Wiley & Sons, Inc.
- DA SILVA, D.N. (2003). Adjustments for Survey Unit Nonresponse Under Nonparametric Response Mechanisms. Thèse de doctorat, Iowa State University, Ames, IA.
- DA SILVA, D.N., et OPSOMER, J.D. (2003). A kernel smoothing method to adjust for unit nonresponse in sample surveys. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association [CD-ROM]. Alexandria, VA. Article #00605.
- FULLER, W.A. (1996). *Introduction to Statistical Time Series* (deuxième édition). Wiley.
- FULLER, W.A., et KIM, J.-K. (2003). Hot deck imputation for the response model. Soumis pour publication.
- HANSEN, M.H., MADOW, W.G. et TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*. 78, 776–793.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 47, 663–685.
- ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*. 77, 89–96.
- KALTON, G. (1983). *Compensating for Missing Survey Data*. Institute of Social Research.
- KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*. 12, 1–17.
- KALTON, G., et MALIGALLO, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. Dans *Proceedings of the Bureau of the Census Annual Research Conference*. U.S. Bureau of the Census (Suitland, MD). 409–428.
- KIM, J.-K., et FULLER, W.A. (1999). Jackknife variance estimation after hot deck imputation. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- ALEXANDRIA, VA. 825–830.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*. 54, 139–157.
- LITTLE, R.J.A., et RUBIN, D.B. (2002). *Statistical Analysis With Missing Data*. Wiley. 20.
- OH, H. L., et SCHEUREN, F.J. (1983). Weighting adjustments for unit non-response. Dans *Incomplete data in sample surveys: Theory and bibliographies*, (Eds. W.G. Madow, I. Olkin, et D.B. Rubin). Academic Press New York: London. 2, 143–184.
- SÄRNÄL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SEN, P.K., et SINGER, J.D.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman and Hall Ltd.
- U.S. BUREAU OF THE CENSUS (1963). The Current Population Survey: A report on methodology. Rapport technique No. 7, Washington, DC.

ANNEXE

Calculs des résultats théoriques

Lemme 1. Supposons que les conditions (A1) à (A3) et (R1) à (R2) sont vérifiées. Définissons pour $i_1, i_2, \dots, i_k \in U_v$

$$\Gamma^{i_1, \dots, i_k} = E \left(\prod_{l=1}^k (I_{i_l} R_{i_l} - \pi_{i_l} \phi_{i_l}) \right),$$

où $\phi_i = \varphi(X_i)$. Considérons le Δ_{i_1, \dots, i_k} de (3). Soit A^r le produit cartésien r -aire de l'ensemble A , où r est un nombre entier positif fixe, $A_{k, r, v} = \{(i_1, i_2, \dots, i_r) \in U_v^r : i_1 = i_2 = \dots = i_r = i\}$ et $A_{k, r, v} = \{(i_1, i_2, \dots, i_r) \in U_v^r : i_1, i_2, \dots, i_r \text{ sont deux à deux distinctes}\}$, $k = 2, 3, \dots, r$. Donc, pour $r = 8$,

$$N_v^{n_v-8} \max_{i_1, \dots, i_8 \in A_{k, 8, v}} (|\Gamma^{i_1, \dots, i_8}|, |\Delta_{i_1, \dots, i_8}|) = \begin{cases} O(N_v^{n_v-4}), & \text{si } k=8, \\ O(N_v^{n_v-4}), & \text{si } k=7, \\ O(N_v^{n_v-5}), & \text{si } k=6, \\ O(N_v^{n_v-4}), & \text{si } k=5 \end{cases}$$

Preuve du lemme 1. Voir Da Silva (2003).

Lemme 2. Supposons que les conditions du théorème 3.1 sont vérifiées. Considérons les vecteurs $t_{gv}^i = (t_{1,g}^i, t_{2,g}^i, t_{3,g}^i) = \sum_{U^i} \pi_{i_l}^i(1, Y_l R_l, R_l) I_{i_l}$, et $t_{gv}^i \equiv (t_{1,g}^i, t_{2,g}^i, t_{3,g}^i)$, avec $t_{3,g}^i = \max\{t_{1,g}^i, t_{2,g}^i, t_{3,g}^i\}$. Soit $t_{gv}^i = E(t_{gv}^i)$. Alors, pour tout $g = 1, 2, \dots, G_v$,

$$\frac{1}{N_g^8} (E\|t_{gv}^i - t_{gv}^i\|, E\|t_{gv}^i - t_{gv}^i\|) = O(G_v^{n_v/4}).$$

Preuve du lemme 2 : Voir Da Silva (2003).

Preuve du théorème 3.1 : Considérons la preuve de (5). Soit $a = (a_1, a_2, a') \in \mathbb{R}^3$ et $h : \mathbb{R}^3 \rightarrow \mathbb{R}$, où $h(a) = a_1 a_2 / a_3, a_3 \neq 0$. Définissons

$$\eta_{gv}^{(a)}(a) = h(N_{-1}^g t_{gv}^{(a)}) + \sum_{k=1}^K h^{(k)}(N_{-1}^g t_{gv}^{(a)})(a_k - N_{-1}^g t_{gv}^{(a)}),$$

où $h^{(k)}(a) = \partial h(a) / \partial a_k$, et soit $e_{gv}^{(a)} = h(a) - \eta_{gv}^{(a)}(a)$. Notons que $t_{gv}^{WC} = \sum_{g=1}^G N_g^8 h(N_{-1}^g t_{gv}^{(a)})$, et, donc, en définissant l'estimateur « linéarisé » $t_{gv}^{WC} = \sum_{g=1}^G N_g^8 \eta_{gv}^{(a)}(N_{-1}^g t_{gv}^{(a)})$, nous pouvons écrire

$$\frac{1}{N} (t_{gv}^{WC} - t_{gv}^{WC}) = \bar{e}_v + \eta_v,$$

où

$$\bar{e}_v = \frac{1}{G} \sum_{g=1}^G N_g^8 e_{gv}^{(a)}(N_{-1}^g t_{gv}^{(a)})$$

et

Quand la variable d'intérêt et le mécanisme de réponse ne sont pas corrélés, même l'estimateur ne comptant qu'une seule cellule de pondération (ce qui correspond à un simple ajustement par quotient) est essentiellement sans biais, tandis que les modèles à cellules multiples donnent d'aussi bons résultats. Par contre, si le mécanisme de réponse et la variable d'intérêt sont corrélés, les propriétés du biais de l'estimateur à cellules de pondération dépendent de façon critique du nombre de cellules. Plus précisément, les estimateurs à une seule cellule sont fortement biaisés, mais, même un nombre assez faible de cellules suffit à réduire aussi bien le biais que la variance de l'estimateur. Ce résultat tient pour les relations linéaires ainsi que non linéaires entre le mécanisme de réponse et la variable d'intérêt.

Même un petit nombre de cellules augmente spectaculairement la performance de l'estimateur. Dans l'ensemble, il semble qu'en présence de non-réponse, créer au moins un petit nombre de cellules de pondération fondées sur une variable reliée à la non-réponse fournit une bonne « police d'assurance » contre le biais dû au plan de sondage et l'inefficacité par rapport au plan de sondage. Nous montrons dans le présent article que cet ajustement ne nécessite pas l'hypothèse que la création des cellules est fondée sur la connaissance a priori de groupes à non-réponse constante. L'estimateur à cellules de pondération résultant ne donnera jamais de moins bons résultats que l'estimateur naïf avec un seul ajustement par quotient pour l'échantillon complet et il pourrait en donner de nettement meilleurs.

6. REMERCIEMENTS

Les auteurs remercient Wayne Fuller pour ses nombreux commentaires constructifs durant l'élaboration du présent manuscrit. Ils remercient aussi le rédacteur adjoint et les deux examinateurs de leurs commentaires. La présente étude a été financée par un contrat de sous-traitance entre Westat et la Iowa State University aux termes du contrat n° ED-99-CO-0109 entre Westat et le U.S. Department of Education. Le premier auteur a mentionner avec reconnaissance l'appui du CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brésil, durant ses études de doctorat à la Iowa State University.

augmentation pour les EQM, spécialement pour les cas où la non-réponse est forte (C1, L1). Pour Y_2 (tableau 5), c'est-à-dire la variable corrélée à X , l'augmentation du nombre de cellules améliore les résultats pour tous les mécanismes de réponse, mais l'effet est nettement plus prononcé quand le mécanisme de réponse est également corrélé à la variable d'intérêt. Comme pour le biais relatif, l'utilisation de trois à cinq cellules donne le gain d'efficacité le plus important, tandis que l'estimateur naïf est très inefficace.

L'écart entre les résultats pour les deux variables est servant des résultats de la section 3. Manifestement, les résultats pour Y_2 suivent la théorie asymptotique, en ce sens que l'EQM s'améliore à mesure que le nombre de cellules augmente (aussi longtemps qu'un nombre suffisant d'observations existe dans chaque cellule). Dans le cas de Y_1 , il convient de souligner pour commencer que le biais est négligable relativement à l'écart-type pour toutes les valeurs de C (voir le tableau 2), de sorte que la variation de l'EQM est due presque exclusivement à des différences de variance. Il s'avère que, quand une variable est iid dans la population et que l'échantillonnage est réalisé avec probabilités égales, la variance asymptotique dans le théorème 3.1 est relativement insensible au nombre de cellules. Dans ce cas, l'augmentation de l'EQM est influencée par la variabilité sous-entendue par l'approximation linéaire dans le théorème 3.1, qui augmente avec le nombre de cellules.

La théorie décrite dans le présent article s'applique aux fonctions de réponse qui peuvent avoir une forme lisse arbitraire. Afin d'évaluer les résultats pour des fonctions plus compliquées, nous créons aussi une variable $X_3 = 25 + 95X - 95X^2 + \varepsilon$, où $\varepsilon \sim N(0, 3)$, de sorte que X_3 a une moyenne de 40,9 et une variance de 51,8, et les deux mécanismes de réponse quadratiques supplémentaires

$$\begin{aligned} - \Phi_{Q1}(X) &= 0,17 + 1,96X - 1,96X^2 \\ - \Phi_{Q2}(X) &= 0,50 + 1,80X - 1,80X^2 \end{aligned}$$

Les résultats (non présentés) reflètent de façon générale ceux obtenus pour les variables antérieures. Si le mécanisme de réponse et les variables sont corrélés (la variable linéaire est corrélée au mécanisme de réponse linéaire et la variable quadratique est corrélée aux mécanismes de réponse linéaire et quadratique), un biais important survient, mais il peut être éliminé en augmentant le nombre de cellules. Dans le cas du mécanisme de réponse quadratique et de la variable quadratique, il semble qu'au moins huit cellules soient nécessaires pour éliminer le biais. De la même façon, l'efficacité relative augmente pour tous les mécanismes de réponse, aussi bien pour la variable

5. CONCLUSION

Aux sections précédentes de l'article, nous avons utilisé un estimateur « linéarisé » \tilde{r}^{WC} , comme approximation de l'estimateur à cellules de pondération, puis nous avons calculé les propriétés asymptotiques de cet estimateur. Par conséquent, il est intéressant de comparer les propriétés statistiques des deux estimateurs dans un contexte de simulation. Pour tous les scénarios présentés au tableau 1, nous calculons l'efficacité relative de l'estimateur à cellules de pondération comparativement à l'estimateur linéarisé. Ces efficacités relatives s'approchent toutes de 1,00, l'écart le plus important étant une valeur de 1,08. Donc, les propriétés statistiques de l'estimateur linéarisé semblent être une bonne approximation de celles de l'estimateur à cellules de pondération.

Nous avons montré que l'estimateur à cellules de pondération, qui correspond aussi à l'estimateur FEF1 proposé par Kim et Fuller (1999), est convergent par rapport au plan d'échantillonnage et un modèle de réponse non paramétrique. Ce modèle ne nécessite pas la spécification correcte des cellules à probabilités de réponse homogènes, à condition qu'on puisse identifier une variable reliée à la probabilité de réponse.

Les propriétés statistiques de l'estimateur dépendent du nombre de cellules utilisées dans l'estimation, mais la relation est assez complexe. Asymptotiquement, il semble qu'il y ait un compromis entre la validité de l'approximation de l'estimateur à cellules de pondération par un estimateur linéarisé, qui nécessite un petit nombre de cellules, et l'erreur quadratique moyenne de cet estimateur linéarisé, qui est réduite si l'on utilise un grand nombre de cellules. Bien qu'ils aident à comprendre le comportement asymptotique de l'estimateur, ces résultats n'offrent que des lignes directrices limitées pour choisir le nombre de cellules pour une enquête particulière. Cependant, ils montrent que la taille d'échantillon dans les cellules doit être raisonnable pour que les estimateurs à cellules de pondération produisent des inférences fiables, parce que les estimations de la variance se fondent habituellement sur la variance de l'estimateur linéarisé comme approximation de la variance des expériences de simulation montrant que, si la variable d'intérêt et le mécanisme de réponse ne sont pas corrélés, le nombre de cellules n'a virtuellement aucun effet sur le biais de l'estimateur du plan d'échantillonnage.

définie comme étant l'EQM de l'estimateur à cellules de pondération divisée par l'EQM de l'estimateur sans non-réponse,

$$EQMR(\hat{f}_{wc}, \hat{f}_y) = \frac{E(\hat{f}_{wc} - \hat{f}_y)^2}{E(\hat{f}_y - t_y)^2}.$$

Dans ces tableaux, la dernière colonne correspond de nouveau à l'EQM relative de l'estimateur naïf. Notons que, à l'exception des deux cas L1 pour la variable Y_2 , les tableaux 4 et 5 sont réellement des tableaux de variance,

puisque le biais est faible.

Tableau 4
Erreur quadratique moyenne relative de l'estimateur à cellules de pondération relativement à l'estimateur sans non-réponse pour Y_1

		Nombre de cellules					Taille de l'échantillon	Mécanisme de réponse	Estimateur naïf
		2	3	5	8				

200	C1	2,02	2,13	2,11	2,21	2,08	500	C1	2,02	2,13	2,11	2,21	2,08
	C2	1,25	1,31	1,29	1,28	1,28		L1	2,34	2,32	2,61	2,70	2,08
	L1	2,34	2,32	2,61	2,70	2,08		L2	1,30	1,29	1,31	1,28	1,28
	C1	2,25	2,21	2,19	2,31	2,23		C2	1,30	1,32	1,34	1,29	1,30
	L1	2,55	2,57	2,62	2,70	2,22		L2	1,32	1,35	1,33	1,34	1,31

Tableau 5
Erreur quadratique moyenne relative de l'estimateur à cellules de pondération relativement à l'estimateur sans non-réponse pour Y_2

		Nombre de cellules					Taille de l'échantillon	Mécanisme de réponse	Estimateur naïf
		2	3	5	8				

200	C1	1,33	1,17	1,10	1,07	2,07	500	C1	1,09	1,05	1,02	1,26	2,07
	C2	1,09	1,05	1,02	1,02	1,26		L1	3,14	1,57	1,16	1,12	26,32
	L1	3,14	1,57	1,16	1,12	26,32		L2	1,23	1,07	1,03	1,01	3,57
	C1	1,35	1,19	1,10	1,09	2,22		C2	1,09	1,05	1,03	1,13	69,75
	L1	6,60	2,30	1,23	1,13	69,75		L2	1,50	1,14	1,04	1,02	7,83

Pour Y_1 (tableau 4), c'est-à-dire la variable non corrélée à X , le nombre de cellules a assez peu d'effets sur l'erreur quadratique moyenne relative, les résultats étant d'environ 2,3 pour un taux de réponse de 50 % et d'environ 1,3 pour un taux de réponse de 80 %. Toutefois, nous observons une

comparables si le mécanisme de réponse est uniforme (C1, C2). Cependant, quand la probabilité de réponse est une fonction linéaire de X (L1, L2), l'estimateur naïf devient sévèrement biaisé. Ce biais relatif diminue à mesure que le nombre de cellules augmente, et de trois à cinq cellules semblent suffisantes pour éliminer la plupart du biais. Cette observation corrobore celle de Cochran (1968) dans le contexte de la réduction du biais pour des études par observation.

Tableau 2
Biais relatif de l'estimateur à cellules de pondération et de l'estimateur naïf pour la moyenne de Y_1

		Nombre de cellules					Taille de l'échantillon	Mécanisme de réponse	Estimateur naïf
		2	3	5	8				

200	C1	0,00	0,00	0,01	0,01	0,00	500	C1	0,00	0,00	0,00	0,00	0,00
	C2	0,01	0,02	0,00	0,00	0,00		L1	0,05	0,02	0,00	0,00	0,01
	L1	0,00	0,03	0,00	0,00	0,00		L2	0,01	0,01	0,00	0,00	0,01
	C1	0,00	0,00	0,04	0,00	0,00		C2	0,01	0,02	0,00	0,00	0,00
	L1	0,00	0,00	0,00	0,00	0,00		L2	0,01	0,01	0,00	0,00	0,01

Tableau 3
Biais relatif de l'estimateur à cellules de pondération et de l'estimateur naïf pour la moyenne de Y_2

		Nombre de cellules					Taille de l'échantillon	Mécanisme de réponse	Estimateur naïf
		2	3	5	8				

200	C1	0,01	0,00	0,00	0,02	-0,01	500	C1	0,01	0,01	0,00	0,00	0,00
	C2	0,00	0,00	0,02	0,01	0,00		L1	1,16	0,59	0,22	0,07	3,57
	L1	1,16	0,59	0,22	0,07	3,57		L2	0,36	0,18	0,06	0,03	1,36
	C1	0,01	0,00	0,00	0,02	-0,01		C2	0,02	0,00	0,00	0,00	-0,01
	L1	1,98	0,96	0,32	0,15	5,84		L2	0,61	0,29	0,09	0,02	2,26

Par conséquent, quand la variable d'intérêt est entièrement indépendante du mécanisme de réponse, comme cela est le cas de Y_1 sous les mécanismes considérés et de Y_2 sous le mécanisme de réponse uniforme, le biais ne dépend pas du nombre de cellules. Quand la variable d'intérêt et le mécanisme de réponse sont reliés, de multiples cellules sont nécessaires pour éliminer le biais.

Nous présentons aux tableaux 4 et 5 l'erreur quadratique moyenne relative (EQMR) pour les deux variables d'intérêt,

correspondent à deux populations de valeurs indépendantes. La variable X_1 a été générée comme $N(40, 58)$, tronquée à -3 et à +3 écarts-types, ce qui correspond au cas du « bruit blanc ». La variable X_2 , qui est reliée à X_1 , a été générée au moyen du modèle linéaire $X_2 = 27,12 + 26,06X_1 + e$, où $e \sim N(0, 9)$. La moyenne de population pour les deux variables sont, respectivement, (39,9, 55,3) pour X_1 et (40,0, 63,9) pour X_2 .

Tableau 1
Vue d'ensemble des facteurs dans l'expérience de simulation

Facteur	Variable Y	Mécanisme de réponse $\phi(\cdot)$	Taille de l'échantillon n	Nombre de cellules G
Niveaux	Y_1, Y_2	$C1, C2, L1, L2$	200, 500	2, 3, 5, 8

Les quatre niveaux du mécanisme de réponse contiennent deux scénarios différents concernant les probabilités de réponse : constantes ($C1, C2$) et reliées linéairement à X ($L1, L2$). Les probabilités de réponse sont :

$$-\phi_{C1}(X) = 0,5$$

$$-\phi_{C2}(X) = 0,8$$

$$-\phi_{L1}(X) = 0,20 + 0,60X$$

$$-\phi_{L2}(X) = 0,65 + 0,30X$$

Les niveaux des mécanismes de réponse linéaires sont choisis de sorte que les probabilités moyennes (sur X) soient approximativement égales à 0,5 et à 0,8, respectivement.

Pour une valeur donnée de G , les groupes sont créés en subdivisant l'intervalle de variation de X en G segments égaux et en assignant l'élément i au groupe g si la valeur X_i est comprise dans le g^{e} segment, $i = 1, 2, \dots, N$ et $g = 1, 2, \dots, G$. Les simulations ont été exécutées conformément à un plan d'expérience factoriel entièrement randomisé $2 \times 4 \times 2 \times 4$. Pour chaque combinaison des niveaux des facteurs dans le tableau 1, nous avons généré $B = 5\,000$ réalisations indépendantes du vecteur d'indicateurs de réponses, $\mathbf{R} = (R_1, R_2, \dots, R_N)^T$, conformément au mécanisme de réponse correspondant. Pour chacune de ces réalisations, nous avons sélectionné un échantillon aléatoire simple (sans remise et de taille n), s , à partir de la population globale. Dans chaque échantillon sélectionné, est pondérée (sans remise) la taille d'échantillon, le mécanisme de réponse et le nombre de cellules considérées. Pour la variable X_2 (tableau 3), les résultats sont

$$\frac{RB(\hat{f}_{wc}, \hat{f}_y)}{E(\hat{f}_{wc} - Y)} = \frac{(\text{Var}(\hat{f}_{wc}^{(Y)}))^{1/2}}{Y}$$

Les tableaux 2 et 3 donnent le biais simulé de l'estimateur à cellules de pondération pour les variables X_1 et X_2 sous forme de fraction de l'écart-type. À titre de comparaison, la dernière colonne des tableaux 2 et 3 contient le biais de l'estimateur naïf, y_f . Le biais sous forme de fraction de l'écart-type, appelé ici *biais relatif*,

4.2 Résultats

ne pourrions pas leur discussion ici. Indistinguables de ceux présentés plus loin, si bien que nous pondération (c^* est-à-dire $\hat{f}_{wc}^{(Y)}$) et les résultats sont presque la valeur du dénominateur dans l'ajustement de la voisine, ainsi qu'une version ayant une borne plus faible sur l'estimateur qui fusionne la cellule vide à une seule cellule voisine. Nous avons également appliqué un choix une valeur plus faible pour G ou fusionner des groupes voisins. Nous avons également appliqué un certain nombre de procédures si les groupes contiennent un trop petit nombre d'éléments, comme utiliser un certain nombre de procédures, on pourrait à considérer dans les simulations. En pratique, on pourrait inférieure raisonnable du nombre d'observations par cellule chaque cellule, nous estimons qu'il s'agit d'une borne inférieure raisonnable du nombre d'observations par cellule. L'estimateur s'appuie sur l'estimation par le quotient dans le nombre qui sera réduit par la non-réponse. Puisque la cellule contiendra environ 25 éléments échantillonnés, pour $n = 200$ et $G = 8$, nous prévoyons que chaque résultat des simulations est négligeable.

En principe, cette procédure pourrait aboutir à un groupe nombre d'échantillons écartés est très faible et l'effet sur les résultats des simulations est négligeable. Pour $n = 200$ et $G = 8$, nous prévoyons que chaque cellule contiendra environ 25 éléments échantillonnés, pour $n = 200$ et $G = 8$, nous prévoyons que chaque résultat des simulations est négligeable. Pour $n = 200$ et $G = 8$, nous prévoyons que chaque

Pour Y_1 (tableau 2), le biais relatif de l'estimateur à cellules de pondération est faible et comparable à celui de l'estimateur naïf, quels que soient la taille d'échantillon, le mécanisme de réponse et le nombre de cellules considérées. Pour la variable X_2 (tableau 3), les résultats sont

et

$$\text{Var} \left(\frac{\tilde{f}_{WC}^v}{N^v} \right) = O \left(\frac{1}{n^v} \right) + O \left(\frac{1}{n^v} \right).$$

Remarque 3. Le théorème 3.2 montre que le biais et la variance asymptotiques de l'estimateur à cellules de pondération \tilde{f}_{WC}^v deviennent tous deux plus petits à mesure que le nombre de groupes G_v augmente. Une explication intuitive est que l'approximation de la fonction $\phi_i = \phi(X_i^v)$ par la fonction en escalier $\phi_i^* = \phi_i^\delta$ s'améliore à mesure que le nombre de cellules augmente. La variance asymptotique comprend un terme qui est indépendant de G_v . Cette « variance résiduelle » est due à la variabilité inhérente au plan d'échantillonnage et au mécanisme de réponse, et ne peut être réduite en modifiant G_v .

Remarque 4. Comme nous le mentionnons à la remarque 1, pour construire une bonne approximation linéaire \tilde{f}_{WC}^v , il faut que G_v soit petit, mais le théorème 3.2 énonce que l'EQM de \tilde{f}_{WC}^v est minimisée en choisissant G_v aussi grand que possible. Regroupées, ces deux conditions peuvent être interprétées comme signifiant que, dès que la taille d'échantillon dans chaque cellule est suffisamment grande pour obtenir un estimateur par le quotient « valide » pour la probabilité moyenne de réponse de la cellule ϕ_i^* , il est préférable d'augmenter le nombre de cellules que d'augmenter la taille de l'échantillon par cellule. L'expérience de simulation discutée à la section 4 nous permettra d'explorer plus en détail cette recommandation.

Le corollaire qui suit découle directement du corollaire 3.1, du théorème 3.2 et de l'égaleité de Chebyshev, et établit la convergence de l'estimateur à cellules de pondération sous le mécanisme de réponse non paramétrique.

Corollaire 3.2. Sous les conditions du théorème 3.2, \tilde{f}_{WC}^v est un estimateur convergent pour t_y , en ce sens que, pour tout $\epsilon > 0$,

$$\Pr \left(\left| \frac{\tilde{f}_{WC}^v - t_y}{N^v} \right| > \epsilon \right) \rightarrow 0, \quad v \rightarrow \infty.$$

Remarque 5. Comme le montre le corollaire 3.2, à condition que l'on puisse trouver une variable X suffisamment reliée à la non-réponse, au sens des hypothèses (R1) à (R3), la construction de cellules de pondération ne nécessite pas la connaissance de cellules à probabilités de réponse homogènes afin de créer un estimateur convergent. Cependant, comme nous en discutons aux remarques 1 et 4, le choix du nombre de cellules continue d'avoir un effet sur les propriétés de l'estimateur.

Remarque 6. L'hypothèse (R3) peut facilement être relâchée pour permettre l'existence d'un petit nombre de points de discontinuité dans $\phi(\cdot)$ ainsi que dans sa dérivée première. Un « petit » nombre peut signifier que le nombre est soit fixe quand $v \rightarrow \infty$, ou qu'il augmente moins rapidement que G_v . Cela permettrait de tenir compte de situations telles que les plans d'échantillonnage stratifiés ou la présence de domaines à l'intérieur de U_v . La présente théorie peut être étendue directement à ces situations si les valeurs de la variable X chevauchent les diverses strates ou domaines.

4. EXPÉRIENCE DE SIMULATION

4.1 Description de l'expérience

Afin d'étudier les conséquences pratiques des résultats de la section 3, nous avons réalisé une expérience de Monte Carlo sur une population fixe de $N = 3\,000$ unités. Nous considérons le cas d'une covariable, X , dont les valeurs de population sont générées comme suit :

$$X_1, X_2, \dots, X_N \sim \text{i.i.d. } U(0, 1),$$

et deux variables d'intérêt distinctes, X_1 et X_2 . Nous voulons évaluer les effets de 1) la relation (modèle) entre X et X_2 , 2) le mécanisme de réponse $\phi(X)$, 3) la taille d'échantillon n et 4) le nombre de cellules G sur le biais et sur l'erreur quadratique moyenne de l'estimateur \tilde{f}_{WC}^v . Puisque nos résultats théoriques s'appuient sur l'approximation de f_{WC} (ou f_{WC}^v) par un estimateur linéarisé \tilde{f}_{WC}^v , nous comparons aussi le comportement de \tilde{f}_{WC}^v/N^v et de \tilde{f}_{WC}^v comme estimateurs de la moyenne de population, $X^v = N^v \cdot \sum_{i=1}^v U_i$. Enfin, nous comparons \tilde{f}_{WC}^v/N^v à l'estimateur « naïf » de la moyenne, qui est défini pour la variable X comme étant :

$$\bar{y}^v = \frac{\sum_{i \in s_y} w_i^v X_i}{\sum_{i \in s_y} w_i^v},$$

correspondant à un rajustement par quotient de l'échantillon de répondants sur l'échantillon original. Cet estimateur est approprié sous l'hypothèse d'un mécanisme de réponse uniforme ou, pour utiliser la terminologie de Little et Rubin (2002, chapitre 1), quand les observations sont *missing completely at random*. Il convient de souligner que \bar{y}^v équivaut à l'estimateur à cellules de pondération ne comptant qu'une seule cellule.

Les niveaux des quatre facteurs utilisés dans l'expérience sont donnés au tableau 1. Les « niveaux » de la variable X

$$\text{Var} \left(\tilde{t}_{WC}^v \right) = \frac{N_v}{1} \sum_{g_v} \sum_{i=1}^{N_g} \left[\sum_{i'=1}^{N_g} \sum_{j=1}^{N_g} \Delta_{ij} \tilde{Y}_{ig}^v \tilde{Y}_{jg}^v \right] + \frac{1}{N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \pi_{ij}^2 \sum_{i'=1}^{N_g} \sum_{j'=1}^{N_g} \left(\tilde{Y}_{ig}^v - \tilde{Y}_{jg}^v \right)^2, \quad (7)$$

où

$$\phi_g = \frac{1}{N} \sum_{i=1}^{N_g} \phi_{i,g}, \quad \tilde{Y}_g = \frac{1}{N} \sum_{i=1}^{N_g} \tilde{Y}_{ig}^v, \quad \tilde{Y}_g^v = \frac{1}{N} \sum_{i=1}^{N_g} \tilde{Y}_{ig}^v$$

et

$$\tilde{Y}_{ig}^v = \frac{\pi_{ij}^2 \phi_{ij}^g}{\phi_{ij}^g + \tilde{Y}_{ig}^v}, \quad \forall i \in U_g \text{ et } \forall g = 1, 2, \dots, G_v.$$

Remarque 1. L'équivalence asymptotique entre \tilde{t}_{WC}^v et

\tilde{t}_{WC}^v dépend du nombre de groupes G_v , la vitesse de convergence étant plus rapide lorsque G_v augmente plus lentement. Ce résultat repose sur le raisonnement intuitif selon lequel la validité de l'approximation linéaire dépend de la justesse de l'estimation des rajustements réels de réponse par les quotients de cellule ϕ_g^* au moyen des estimateurs fondés sur l'échantillon $\sum_{i=1}^{N_g} w_i^g$. Puisque les quotients de cellule deviennent de meilleurs estimateurs à mesure que la taille de l'échantillon augmente, on peut soutenir que G_v devrait être choisi petit, ce qui correspond à la pratique courante dans les applications d'estimation par cellule de pondération. Cependant, comme nous le montrons plus loin, les propriétés de l'EQM de \tilde{t}_{WC}^v sous le mécanisme de réponse non paramétrique s'améliorent à mesure que G_v grandit. Nous discutons plus en détail de la sélection du nombre de groupes après l'exposé du théorème 3.2 présenté plus bas et à la section 4.

Remarque 2. Les résultats du théorème 3.1 dépendent des

groupes de population $U_g^v, g = 1, \dots, G_v$ et des $\phi_{ij}^g, i \in U_g^v$, mais ne s'appuient pas sur le fait que la probabilité de réponse est une fonction lisse de la variable auxiliaire X . Par conséquent, nous pouvons utiliser les expressions explicites du biais et de la variance asymptotiques pour calculer les résultats pour d'autres mécanismes de réponse qui obéissent à (R1) et (R2). Plus précisément, les résultats pour le modèle des *groupes à réponse homogène* (voir Sæthdal et coll. 1992, page 577) découlent directement du théorème 3.1. Il s'agit aussi du modèle étudié par Fuller et Kim (2003). Sous ce modèle, nous supposons que $\phi_{ij}^g \equiv \phi_g^*$ pour tout $i \in U_g^v, g = 1, \dots, G_v$, et nous pouvons montrer facilement que le biais de \tilde{t}_{WC}^v est 0 et que sa variance est

$$\text{E} \left(\tilde{t}_{WC}^v \right) = \tilde{t}_{WC}^v - Y_v^* = O \left(\frac{1}{N_v} \right)$$

*théorème 3.1 tiennent. Alors,***Théorème 3.2.** Supposons que (R3) et les conditions du

pour le biais et la variance et est prouvé à l'annexe. \tilde{t}_{WC}^v . Le théorème suivant donne les taux asymptotiques de groupes G_v sur le biais et la variance asymptotiques de par (R1) à (R3), nous pouvons décrire l'effet du nombre de

Sous le mécanisme de réponse non paramétrique décrit résultat est assez général. estimateur par extension classique de la forme (1), ce même de \tilde{t}_{WC}^v . Puisque \tilde{t}_{WC}^v peut s'écrire comme un \tilde{t}_{WC}^v atteint asymptotiquement la normalité, alors il en est de

$$\left[\text{Var} \left(\tilde{t}_{WC}^v \right) \right]^{-1/2} \left[\tilde{t}_{WC}^v - Y_v^* \right] \xrightarrow{L} N(0, 1).$$

alors

$$V \equiv \lim_{v \rightarrow \infty} n_v \text{Var}(\tilde{t}_{WC}^v) / N_v \in (0, \infty),$$

où B_v correspond au biais de \tilde{t}_{WC}^v donné dans le

$$n_{1/2} \left(\tilde{t}_{WC}^v - Y_v^* - B_v \right) \xrightarrow{L} N(0, V),$$

tel que

Corollaire 3.1. Sous les conditions du théorème 3.1 avec $\gamma < 1/2$ dans (A7), pour tout plan d'échantillonnage $p_v(\cdot)$

preuve figure à l'annexe.

Le corollaire qui suit découle directement du théorème 3.1 et de Fuller (1996, théorème 5.2.1). Une l'intérieur des cellules.

non-réponse sous un mécanisme de réponse homogène à terme représente l'inflation de la variance causée par la variance de l'estimateur sans non-réponse et le deuxième Le premier terme de l'expression de la variance est la

$$\text{Var} \left(\tilde{t}_{WC}^v \right) = \text{Var} \left(\tilde{t}_{WC}^v \right) + \frac{1}{N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \pi_{ij}^2 \left(\tilde{Y}_{ig}^v - \tilde{Y}_{jg}^v \right)^2.$$

Fuller et Kim (2003) donnent la loi limite de l'estimateur FEF sous l'hypothèse que les probabilités de réponse sont constantes à l'intérieur des cellules. Nous étudions le cas où la probabilité de réponse est fonction lisse d'une variable auxiliaire et où le nombre de cellules peut varier. Soit $\mathbf{R}^v = (R_1^v, R_2^v, \dots, R_{N^v}^v)^T$ le vecteur d'indicateurs de réponse pour la v^e population. Nous supposons que la loi de \mathbf{R}^v satisfait aux hypothèses de mécanisme de réponse non paramétrique énoncées comme suit :

(R1) $R_1^v, R_2^v, \dots, R_{N^v}^v$ sont des variables aléatoires indépendantes,

(R2) $\Pr \{R_i^v = 1 | \mathbf{1}^{v^v}, X^v\} = \phi_i^v, \forall i \in U^v,$

(R3) $\phi_i^v = \phi(X_i^v), \forall i \in U^v$, où $\phi(\cdot)$ est dérivable avec une dérivée première bornée et les $X_i^v \in [x_m^v, x_M^v]$, avec x_m^v, x_M^v constantes fixes et $x_m^v < x_M^v$.

Les autres hypothèses sont des conditions techniques que nous utilisons largement dans les preuves. Nous supposons qu'il existe des constantes positives $\lambda_1, \lambda_2, \dots, \lambda_\gamma$ telles que :

(A1) $\lambda_1 < N^v \pi_i^v < \lambda_2 < \infty, \forall i \in U^v$, et $N^v N^{v-1} \pi_i^v \rightarrow \pi \in (0, 1)$, quand $v \rightarrow \infty$;

(A2) Pour des $i_1, \dots, i_K \in U^v, K = 2, 3, \dots, 8$ distincts,

$$|\Delta_{i_1, \dots, i_K}^v| \leq \begin{cases} \left(\prod_{k=1}^{K-1} (N^v - k + 1) \right)^{-1} N^{v/K2} \lambda_{i_K}^v, & \text{si } K \text{ est pair} \\ \left(\prod_{k=1}^{K-1} (N^v - k + 1) \right)^{-1} N^{v/(K-1)2} \lambda_{i_K}^v, & \text{si } K \text{ est impair} \end{cases}$$

(A3) $\lim_{v \rightarrow \infty} \frac{1}{N^v} \sum_{i \in U^v} \phi_i^v = \phi_\pi^v, \forall \pi = 1, 2, \dots, G^v$ et $v \geq 1$;

(A4) $\max_{i \in U^v} |X_i^v| \leq \lambda_\gamma$;

(A5) $\lambda_6 < \min_{i \in U^v} \phi_i^v \leq 1$;

(A6) $\lambda_\gamma G^{v-1} \leq N^v \leq \lambda_\gamma G^{v-1}, \forall \pi = 1, 2, \dots, G^v$;

(A7) $1 \leq G^v \leq n_\gamma^v \lambda_\gamma$, avec $0 \leq \gamma \leq 1/2$.

Les hypothèses (A1) et (A2) impliquent qu'asymptotiquement, le plan d'échantillonnage « se comporte bien », en ce sens que les moments des indicateurs d'inclusion dans l'échantillon sont du même ordre de grandeur que ceux observés pour l'échantillon aléatoire simple sans remise. Il s'agit d'une hypothèse courante de la théorie asymptotique en population finie. (A1) exige aussi que la traction d'échantillonnage converge vers une constante dans l'intervalle $(0, 1)$. L'hypothèse (A4) de limitabilité, c'est-à-dire la capacité d'admettre des bornes, concernant

et

$$(6) \quad E \left(\tilde{t}_{WC}^{N^v} \right) - \bar{Y}^v = \frac{1}{G^v} \sum_{g=1}^{G^v} \sum_{i \in U^g} \left(\frac{\phi_i^g}{\phi_i^g - \bar{\phi}^g} \right) \left(X_i^g - \bar{Y}^g \right)$$

Le biais et la variance de $\tilde{t}_{WC}^{N^v}$ sont donnés par

$$(5) \quad \frac{1}{N^v} (t_{WC}^{N^v} - \bar{t}_{WC}^{N^v}) = O^p(G^v n_\gamma^{v-1}).$$

linéarisée $\tilde{t}_{WC}^{N^v}$, en ce sens que

asymptotiquement équivalent à une variable aléatoire hypothèses (A1) à (A7) tiennent. Alors, l'estimateur $\tilde{t}_{WC}^{N^v}$ est conditions (R1) et (R2). Enfin, supposons que les $p(\cdot)$, et que le mécanisme de réponse satisfait aux partir de U^v conformément au plan d'échantillonnage un échantillon probabiliste de taille fixe $n_\gamma^v (n_\gamma^v \geq n^{v-1})$ à $\{U^v : v \geq 1\}$. Supposons que, pour chaque v , on sélectionne

Théorème 3.1. Considérons la série de populations

résultats. La preuve figure à l'annexe.

Le théorème qui suit énonce formellement nos premiers par exemple, Särndal et coll. (1992, chapitre 5).

ration. Pour une description de cette approche, consulter, de la loi asymptotique de l'estimateur à cellules de pondération étant les propriétés asymptotiques de l'estimateur à cellules de pondération ou, plus précisément, les propriétés moyenne de l'estimateur linéarisé et nous les considérons nous calculons les propriétés de l'erreur quadratique ration non linéaire et une approximation « linéarisée ». Puis, l'ence asymptotique entre l'estimateur à cellules de pondération finie. Premièrement, nous montrons l'équivalence habituellement adoptée pour l'étude des estimateurs en priétés de l'estimateur à cellules de pondération suit celle L'approche que nous utilisons dans l'étude des pro-

3.2 Principaux résultats

comparativement à la taille de l'échantillon.

nombre total de cellules n'augmente pas « trop rapidement » croissance soit le même pour toutes les cellules et que le sur les cellules de pondération exigent que le taux de pourrait avoir $\phi_i^v = 0$. Enfin, les hypothèses (A6) et (A7) cellule et (A5) exclut la situation où certaines unités existe pour la probabilité moyenne de réponse dans une des mécanismes de réponse : (A3) assure que la limite régularité sont nécessaires pour éviter la dégénérescence bornes. De même, certaines conditions techniques de relâchée au besoin, en postulant l'existence de moments de certains théorèmes utilisés dans l'article et pourrait être les observations simplifiera considérablement les preuves

cet estimateur. Pour éviter ce problème, nous nous limitons dans le présent article à un mécanisme de réponse assez général. Plus précisément, nous supposons que les R_i sont des variables indépendantes de Bernoulli telles que

$$P\{R_i = 1 | I, X\} = \varphi_i, \quad 0 < \varphi_i \leq 1, \forall i \in U,$$

et que les φ_i peuvent s'écrire $\varphi_i = \varphi(X_i)$, où $\varphi(\cdot)$ est une fonction continue et dérivable, mais par ailleurs non spécifiée, de X_i . Notons que ceci inclut le mécanisme de réponse uniforme, où $\varphi_i \equiv \varphi$ pour tout $i \in U$, comme cas particulier.

Si certains éléments sélectionnés ne répondent pas, l'estimateur (1) ne peut plus être calculé et il est nécessaire d'utiliser un estimateur qui inclut une correction pour la non-réponse. Dans le présent article, nous utilisons l'estimateur à cellules de pondération à cette fin. Par souci de simplicité, nous décrivons la situation où les X_i ainsi que les X_i^* sont des variables univariées, mais l'approche peut être généralisée au cas multidimensionnel. Soit $s_r = s \cap r$ le sous-ensemble d'éléments sélectionnés qui répondent effectivement à l'enquête.

Représentons par $U_g, g = 1, \dots, G$, les G groupes obtenus en subdivisant la population en groupes d'après les valeurs de la variable auxiliaire connue X . Des exécutions partielles pourraient consister à générer des groupes de taille égale ou à subdiviser l'intervalle de variation de X en intervalles de longueur égale. Pour l'instant, nous ne spécifions pas l'exécution, et nous énonçons certaines hypothèses générales au sujet de G et de la taille des groupes à la section suivante. Souignons que nous considérons que les groupes sont fixes par rapport au plan d'échantillonnage et au mécanisme de réponse, ce qui exclut la situation où les groupes sont formés d'après les valeurs d'échantillon observées $\{X_i : i \in s\}$. Nous adoptons cette approche, qui est notamment semblable à celle de Särndal et coll. (1992) et de Kim et Fuller (1999), avant tout pour simplifier les calculs théoriques.

Soit $s_g = s \cap U_g$ la partie de l'échantillon qui tombe dans le groupe g , et définissons simultanément $s_{r,g} = s_r \cap U_g$. L'estimateur à cellules de pondération est défini comme étant

$$\hat{t}_{WC} = \sum_g \left(\frac{\sum_{i \in s_{r,g}} w_i}{\sum_{i \in s_g} w_i} \right) \sum_{i \in s_{r,g}} w_i X_i. \quad (2)$$

Partant de cette expression, il est facile de voir que, dans chaque groupe, l'estimateur du total de groupe est corrigé par application d'un quotient égal à l'inverse de la proportion pondérée de répondants dans la cellule. Cet estimateur est également l'estimateur FEHT de Kim et Fuller (1999). Nous étudions ses propriétés à la section suivante.

3. PROPRIÉTÉS SOUS QUASI-RANDOMISATION

3.1 Cadre asymptotique et hypothèses

Nous étudions les propriétés sous quasi-randomisation de l'estimateur à cellules de pondération dans le contexte asymptotique en population finie habituel, dans lequel la population U est traitée comme un élément d'une série croissante U_1, U_2, \dots, U_v avec $v \rightarrow \infty$, à laquelle est associée une série correspondante de plans d'échantillonnage $p_v(\cdot)$ (voir Isaki et Fuller (1982) pour un des premiers exemples de ce cadre de travail). Soit N_v la taille de la v^e population avec $N_v > N_{v-1}$, soit $X_v = (X_1^v, X_2^v, \dots, X_{N_v}^v)^T$ l'ensemble de valeurs de la caractéristique d'intérêt Y , associée à U_v , et, par conséquent, $X_v = (X_1^v, X_2^v, \dots, X_{N_v}^v)^T$. Nous supposons que X_v est connu. Pour chaque v , un échantillon de taille n_v ($n_v \geq n_{v-1}$) est sélectionné à partir de U_v , conformément à un plan d'échantillonnage $p_v(\cdot)$. Comme précédemment, soit le vecteur d'indicateurs d'inclusion $I^v = (I_1^v, I_2^v, \dots, I_{N_v}^v)^T$ dans l'échantillon correspondant. Nous représentons le moment central de K^e ordre des indicateurs d'inclusion dans l'échantillon $I_{i_1^v, \dots, i_{K^e}}^v$ par

$$\Delta_{i_1^v, \dots, i_{K^e}}^v = E \left(\prod_{k=1}^{K^e} (I_{i_k^v}^v - \pi_{i_k^v}^v) \right). \quad (3)$$

Nous supposons que U_v peut être subdivisée en G_v ($G_v \geq G_{v-1}$) groupes mutuellement exclusifs et exhaustifs, $U_{g_v}, g_v = 1, \dots, G_v$. Pour construire ces groupes, nous trions la population en fonction des valeurs de X et nous la subdivisons en G_v groupes. Nous supposons qu'il existe au moins G_v valeurs distinctes parmi les éléments de X_v . Soit N_g^v le nombre d'éléments dans U_{g_v} . Comme nous l'avons mentionné à la section précédente, nous traitons les groupes comme étant fixes par rapport à la population. Le problème que pose cette approche est qu'en général, la probabilité d'obtenir un groupe ne contenant aucun répondant n'est pas nulle. Nous résolvons ce problème en ajoutant une petite constante dans le dénominateur de chaque groupe, ou

$$\hat{t}_{WC}^v = \sum_{g_v} \left(\frac{\max \left(\sum_{i \in s_{r,g_v}} w_i^v, N_g^v G_v n_v^{v-1} \right)}{\sum_{i \in s_{r,g_v}} w_i^v} \right) \sum_{i \in s_{r,g_v}} w_i^v X_i^v. \quad (4)$$

Donc, la différence entre \hat{t}_{WC}^v et \hat{t}_{WC}^v dans (2) est asymptotiquement négligeable. Cette façon de procéder est semblable à ce qui se fait fréquemment en pratique pour éviter des poids exagérément grands dans l'estimation par quotient.

connu. Pour chaque élément i de U , soit $X_i^1 = (X_{1,i}^1, X_{2,i}^1, \dots, X_{p,i}^1)$ le vecteur connexe de valeurs de p caractéristiques d'intérêt, $X_1^1, X_2^1, \dots, X_p^1$. De même, soit $X_i^t = (X_{1,i}^t, X_{2,i}^t, \dots, X_{q,i}^t)$, le vecteur de valeurs de q variables auxiliaires, $X_1^t, X_2^t, \dots, X_q^t$, correspondant à la t^e unité, $t \in U$. Nous supposons que X_i^t est connu $\forall i \in U$. Si $p = 1$, nous désignons X_i^1 par X_i^t et, pour $q = 1$, nous utilisons X_i^q pour désigner X_i^t . Représentons par s un échantillon tiré de U selon un plan d'échantillonnage $p(\cdot)$. Ce plan d'échantillonnage $p(\cdot)$ est choisi par l'échantillonneur et peut être fondé sur l'information disponible dans les $X_i^t, i \in U$.

Le but de l'enquête par sondage est d'estimer des quantités de population inconnues, comme la moyenne ou le total de population, ou une fonction de ces quantités. Pour simplifier la présentation, nous nous concentrons sur l'estimation du total de population des X_i^t .

$$t_y = \sum_{i \in U} X_i^t.$$

Si n y a pas de non-réponse, cette quantité est estimée au moyen d'un estimateur fondé sur l'échantillon de la forme

$$(1) \quad t_y = \sum_{i \in s} w_i^t X_i^t = \sum_{i \in s} w_i^t X_i^1 X_i^t I_i^t$$

où les $w_i^t, i \in s$, sont les poids d'échantillonnage et I_i^t est un indicateur précisant si la t^e unité est comprise ou non dans l'échantillon. Dans le présent article, nous supposons que les poids d'échantillonnage sont égaux à l'inverse des probabilités d'inclusion, soit $w_i^t = \pi_i^{-1}$, avec $\pi_i = P(i \in s)$, si bien que l'estimateur (1) est l'estimateur d'Horvitz-Thompson classique (Horvitz et Thompson 1952). En outre, nous représentons par $I = (I_1^t, I_2^t, \dots, I_N^t)$ le vecteur d'indicateurs d'inclusion pour la population.

Dans le contexte de la non-réponse, il est commode de supposer que chaque unité de la population est soit un *répondant*, soit un *non-répondant* pour la variable d'intérêt X . Considérons le vecteur $R = (R_1, R_2, \dots, R_N)^T$, où R_i indique si la i^e unité est ou non un répondant. La loi de R est appelée *mécanisme de réponse*. Par analogie avec la définition de l'échantillon s , nous utilisons $r \subseteq U$ pour désigner l'ensemble (réalisé) de répondants dans la population, autrement dit les éléments pour lesquels $R_i = 1$. Puisque la loi de r et de R est généralement inconnue et peut, théoriquement, dépendre de la valeur réalisée de I , ainsi que de X , nous devons représenter le mécanisme de réponse par un modèle hypothétique. Si nous utilisons ce modèle hypothétique pour élaborer un estimateur pour une quantité de population, les propriétés de cet estimateur deviennent dépendantes du modèle de réponse. Par conséquent, un modèle spécifique incorrectement pour R peut induire un biais significatif et difficile à mesurer dans l'estimateur, ainsi que dans les mesures de précision connexes de

mécanisme de réponse non paramétrique sont fournies à la section suivante). La connaissance d'une telle variable pourrait être utilisée pour construire des corrections plus élaborées pour la non-réponse, comme la pondération d'après la *propension à répondre* (Cassel, Särndal et Wretman (1983), Little (1986) et Da Silva et Opsomer (2003)) ou la stratification a posteriori, mais nous limiterons ici l'utilisation de cette variable auxiliaire à la subdivision de la population en cellules de pondération. L'objectif principal de cette approche est d'étudier la robustesse de l'estimateur à cellules de pondération bien connu aux erreurs de spécification du modèle, en particulier, l'effet du nombre de cellules. Donc, contrairement à l'approche des auteurs susmentionnés, nous utilisons les cellules de pondération de façon pratique pour construire un estimateur applicable à des données d'enquête, mais il n'est pas supposé qu'elles font partie du cadre statistique. Cette approche est comparable à la notion de « ajustement par sous-classification » proposée par Cochran (1968) pour éliminer le biais dû à une covariable continue dans les études par observation.

Nous étudions les propriétés de l'estimateur sous *quasi-randémisation*, un terme utilisé par Oh et Scheuren (1983) pour indiquer une inférence fondée conjointement sur le plan d'échantillonnage et sur le mécanisme de réponse. Nous établissons les propriétés asymptotiques de l'estimateur en intégrant la population finie, ainsi que le plan de sondage et le mécanisme de réponse correspondants dans une série de populations et de mécanismes aléatoires de ce genre, comme nous l'expliquerons plus loin. Ce cadre asymptotique ressemble fort à celui préconisé par Hansen, Madlow et Tepping (1983) et utilisé dans Isaki et Fuller (1982), entre autres.

Le reste de l'article est présenté comme suit. À la section 2, nous introduisons la notation et le cadre pour le plan de sondage et le modèle de réponse, et nous discutons de l'estimateur à cellules de pondération. À la section suivante, nous calculons les propriétés asymptotiques de l'estimateur par rapport au plan de sondage. À la section 4, nous présentons les résultats d'une étude en simulation pour examiner le comportement de l'estimateur en pratique, comparer ce comportement à celui prévu par la théorie asymptotique et offrir certaines lignes directrices quant au choix des cellules de pondération.

2. L'ESTIMATEUR À CELLULES DE PONDERATION

Avant de décrire l'estimateur à cellules de pondération, nous introduisons notre cadre de conception de l'enquête et le mécanisme de génération des réponses. Nous considérons une population $U = \{1, 2, \dots, N\}$, où N est fini et

Propriétés de l'estimateur à cellules de pondération sous un mécanisme de réponse non paramétrique

D. NASCIMENTO DA SILVA et JEAN D. OPSOMER¹

RÉSUMÉ

L'estimateur à cellules de pondération corrige la non-réponse totale par subdivision de l'échantillon en groupes homogènes (cellules) et application d'une correction par quotient aux répondants compris dans chaque cellule. Les études antérieures des propriétés statistiques des estimateurs à cellules de pondération se fondaient sur l'hypothèse que ces cellules correspondent à des cellules de population connues dont les caractéristiques sont homogènes. Dans le présent article, nous étudions les propriétés de l'estimateur à cellules de pondération sous un modèle de probabilité de réponse qui ne nécessite pas la spécification correcte de cellules de population homogènes. Nous supposons plutôt que la probabilité de réponse est une fonction lisse, mais par ailleurs non spécifiée, d'une variable auxiliaire connue. Sous ce modèle plus général, nous étudions la robustesse de l'estimateur à cellules de pondération à la spécification incorrecte du modèle. Nous montrons que, même si les cellules de population sont inconnues, l'estimateur est convergent par rapport au plan d'échantillonnage et au modèle de réponse. Nous décrivons l'effet du nombre de cellules de pondération sur les propriétés asymptotiques de l'estimateur. Au moyen d'expériences de simulation, nous explorons les propriétés de population finie de l'estimateur. Pour conclure, nous donnons certaines lignes directrices concernant le choix de la taille des cellules et de leur nombre pour l'application pratique de l'estimation fondée sur des cellules de pondération lorsqu'on ne peut spécifier ces cellules a priori.

MOTS CLÉS : Propriétés asymptotiques en population finie; inférence sous quasi-randomisation; sélection de cellules de pondération.

1. INTRODUCTION

La non-réponse partielle et la non-réponse totale ont lieu dans presque toutes les enquêtes à grande échelle, et des techniques d'estimation appropriées sont nécessaires pour en tenir compte. Tandis que la non-réponse partielle est souvent traitée par imputation, la non-réponse totale est le plus fréquemment corrigée par repondération. La repondération par cellule pour tenir compte de la non-réponse remonte au moins aux années 1950 dans le domaine de Bureau of the Census (1963, page 33), et son usage reste très répandu de nos jours, parce que la méthode est intuitivement séduisante et assez facile à appliquer en pratique. Kalton (1983) et Kasprzyk (1986) ont passé en revue les procédures courantes de pondération. Plusieurs auteurs ont étudié les propriétés de l'estimateur à cellules de pondération sous divers cadres théoriques. Oh et Scheuren (1983) calculent la moyenne et la variance de cet estimateur sous échantillonnage aléatoire simple, conditionnellement à la taille de l'échantillon et au nombre de répondants dans chaque cellule. À cet égard, voir aussi Katon et Maligaig (1991), Samdal, Swensson et Wretman (1992, page 578) utilisent l'expression « groupe à homogenité de réponse » pour les cellules dans lesquelles il est supposé que la non-réponse est constante et calculent les propriétés de l'estimateur à cellules de pondération

réulnant pour des plans de sondage généraux. La méthode d'imputation fractionnaire entièrement efficace [fully efficient fractional imputation (FEFI)] introduite récemment par Kim et Fuller (1999) peut aussi être exprimée sous forme d'un estimateur à cellules de pondération, et ces auteurs calculent les propriétés de modèle de l'estimateur sous l'hypothèse que les variables sont indépendantes et identiquement distribuées (iid) dans chaque cellule. Bien que les hypothèses particulières varient, une trame commune à tous ces résultats est que les cellules de pondération sont spécifiées correctement, en ce sens que, dans chaque cellule, les unités sont entièrement « interchangeables » (la définition précise de ce terme dépend du cadre analytique choisi : probabilités égales de réponse pour l'inférence fondée sur la randomisation, ou observations iid pour l'inférence fondée sur un modèle). Dans la terminologie de Little et Rubin (2002, chapitre 1), il s'agit du cas d'observations *manquant au hasard* [(MAR) pour *missing at random*], où l'information auxiliaire (c'est-à-dire l'appartenance à la cellule dans ce cas-ci) peut être utilisée pour corriger l'inférence pour la non-réponse. Dans le présent article, nous nous écartons de ce cadre analytique. Nous supposons que le mécanisme de réponse dépend d'une variable auxiliaire connue, mais la forme fonctionnelle exacte de cette relation demeure presque entièrement non spécifiée (des précisions sur ce

¹ D. Nascimento Da Silva, Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN 59072-970, Brésil. Courriel : damiao@cce.ufrrn.br; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames IA 50011, États-Unis. Courriel : jopsomer@iastate.edu.

- et
- $$\mathcal{Q}_2 = \left[\sum_{i \in A_2} w_{(1)}' \right]^{-1} (N_1 + N_2 - \hat{N}_1),$$
- Les $W_{(2)}'$ dans $A_1 \cup A_2$, l'union des cellules 1 et 2, maintiennent le nombre total de ménages dans $A_1 \cup A_2$ et fournissent un total estimatif pour la cellule 1 qui est raisonnablement proche du total réel.
- (iii) Les cellules 1 et 2 sont désignées pour le regroupement, $n_1 + n_2 \geq 5$, et $(n_1 + n_2)^{-1} (N_1 + N_2) > B$. Alors, il est nécessaire de pousser le regroupement plus loin. La cellule combinée devient la cellule 1 pour le cas (ii). Pour suivre le regroupement des cellules jusqu'à ce que $(n_1 + n_2 + \dots)^{-1} (N_1 + N_2 + \dots) \leq B$. Le cas (iii) n'a pas été observé pour l'ensemble de données étudié.
- On pourrait répéter la procédure de calcul des poids de façon itérative en substituant $W_{(2)}'$ à $W_{(1)}'$ dans un deuxième cycle. L'exécution d'un deuxième cycle sur les données décrites dans le texte n'a donné lieu à aucune amélioration discernable des estimations.
- ## BIBLIOGRAPHIE
- ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.
- BANKIER, M.D., RATHWELL, S. et MAJKOWSKI, M. (1992). Two Step Generalized Least Squares Estimation in the 1991 Canadian Census. Document de travail, Direction de la méthodologie, Division des méthodes d'enquêtes sociales, Statistique Canada. SSM92-007E.
- BANKIER, M., HOULE, A.M. et LUC, M. (1997). Calibration Estimation in the 1991 and 1996 Canadian Censuses. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 66-75.
- BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- DEVILLE, J., et SÄRDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- FULLER, W.A. (1998). Replication variance estimation for two phase samples. *Statistica Sinica*, 8, 1153-1164.
- FULLER, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- FULLER, W.A., et ISAKI, C.T. (2001). Estimation Using Estimated Coverage in a Census. Présenté à la CAESAR conference, juin, Rome, Italie.
- FULLER, W.A., LAUGHIN, M.M. et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.
- HOGAN, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. *Journal of the American Statistical Association*, 88, 1047-1060.
- HUSAIN, M. (1969). Construction of Regression Weights for Estimation in Sample Surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.
- ISAKI, C.T., TSAY, J.H. et FULLER, W.A. (2000). Estimation des facteurs de correction au recensement. *Techniques d'enquête*, 26, 37-49.
- ISAKI, C.T., IKEDA, M.M., TSAY, J.H. et FULLER, W.A. (2000). An estimation file that incorporates auxiliary information. *Journal of Official Statistics*, 16, 155-172.
- JAYASURIYA, B.R., et VALLANT, R. (1996). Application de l'estimation par régression restreinte dans une enquête-ménage. *Techniques d'enquête*, 22, 127-138.
- LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- RAO, J.N.K., et SINGH, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-65.
- SCHINDLER, E., GRIFFIN, R. et SWAN, C. (1992). Weighting the 1990 Census Sample. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 664-669.

5. CONCLUSION

Nous remercions le rédacteur adjoint et deux examinateurs de leurs commentaires qui nous ont permis d'améliorer considérablement l'article

ANNEXE

Procédures utilisées pour définir les cellules et les poids initiaux $W_i^{(2)}$

Nous utilisons la procédure de l'USCB pour déterminer l'ordre dans lequel les cellules sont combinées (regroupées). Les règles de regroupement des cellules précisent que chaque cellule doit contenir au moins cinq ménages échantillonnés. La procédure qui suit est notre extension des règles de l'USCB pour définir $W_i^{(2)}$.

Considérons deux cellules, nommées cellule 1 et cellule 2.

i) La cellule 1 ne doit pas être regroupée et $n_1^{-1}N_1 \leq B$, où N_1 est le dénombrement de ménages au recensement dans la cellule 1 et n_1 est le nombre d'unités d'échantillonnage pour le questionnaire détaillé dans la cellule 1. La constante B est fournie par l'organisme patron et, dans nos travaux, nous utilisons la valeur 27. Pour le ménage i dans la cellule 1, soit

$$W_i^{(2)} = \max\{1, 2, W_i^{(1)}\}, \quad (A.1)$$

où

$$W_i^{(1)} = \min\{Q_1 W_i^{(0)}, B\},$$

$$Q_1 = \left[\sum_{i \in A_1} W_i^{(0)} \right]^{-1} N_1,$$

et A_1 est l'ensemble d'indices dans la cellule 1. Le nombre 1,2 est une borne inférieure arbitraire choisie de sorte qu'elle soit supérieure à l'unité et inférieure à la valeur minimale de $W_i^{(0)}$, c'est-à-dire deux. Notons que les $W_i^{(2)}$ fournissent des totaux estimés raisonnables pour la cellule 1. Si $n_1^{-1}N_1 > B$, regrouper la cellule 1 avec la cellule 2 comme il est décrit en ii) ci-après.

ii) Les cellules 1 et 2 sont désignées pour le regroupement, $(n_1 + n_2)^{-1}(N_1 + N_2) \leq B$, $n_1 + n_2 \geq 5$, et $n_1^{-1}N_1 > n_2^{-1}N_2$. Alors, pour i dans la cellule 1, $W_i^{(2)}$ est défini par (A.1). Pour i dans la cellule 2,

$$W_i^{(2)} = \max\{1, 2, W_i^{(1)}\},$$

où

$$W_i^{(1)} = \min\{Q_2 W_i^{(0)}, B\},$$

Nous montrons que la méthode de programmation quadratique donne de bons résultats quand on l'applique aux données réelles provenant du questionnaire détaillé de l'USCB. La méthode PQ avec poids unique au niveau du ménage présente plusieurs avantages par rapport à la méthode avec poids distincts de l'USCB. L'utilisation d'un seul ensemble de poids élimine toute confusion quant au poids qu'il convient d'utiliser pour estimer une caractéristique donnée. En outre, les estimations de relations telles que les ratios de caractéristiques individuelles par rapport aux caractéristiques du ménage devraient être moins variables si l'on utilise un seul ensemble de poids pour les deux types de caractéristiques.

Étant donné qu'il est plus facile pour les analystes de ne calculer et n'utiliser qu'un seul ensemble de poids, on ne produirait deux ensembles de poids que si ceux élaborés spécialement pour un type de caractéristique produisent des estimations dont la variance est plus faible pour ce type de caractéristique, ce qui ne semble pas être le cas dans notre exemple. L'ensemble unique de poids PQ produit des résultats favorables pour les caractéristiques du ménage ainsi que celles de la personne comparativement à ceux obtenus en utilisant les poids USCB pour la catégorie particulière.

Le module d'estimation par programmation quadratique module d'estimation par la méthode itérative du quotient (raking ratio) dans le système opérationnel de l'USCB. La méthode PQ permet de produire des poids d'échantillonnage pour le questionnaire détaillé au niveau des ménages dans une situation de rajustement où seuls des contrôles au niveau de la personne sont disponibles.

REMERCIEMENTS

Le présent article présente les résultats de travaux de recherche et d'analyse entrepris par les auteurs. Il a fait l'objet d'un examen plus limité que les publications officielles du U.S. Census Bureau. Les résultats de recherche et les conclusions présentées n'engagent que les auteurs et non pas été évalués par le U.S. Census Bureau. Le rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche réalisés et de favoriser la discussion.

Cette étude a été financée en partie par l'entente de collaboration 43-3AEU-3-80088 entre la Iowa State University, le National Agricultural Statistics Service et le U.S. Bureau of the Census.

niveau de la personne d'après les données de l'enquête postcensitaire de 1990; par contre, aucune estimation au niveau de l'unité de logement n'existe pour cette enquête. Nous donnons à ces estimations le nom d'estimations ACE. Voir Hogan, (1993) et Isaki, Tsay et Fuller (2000). Les estimations pour le SP 1788 ont été calculées par la méthode PQ, en utilisant les estimations ACE comme fonction objectif avec 63 caractéristiques âge-race-sexe-mode d'occupation du logement au niveau de la personne dans le deuxième terme de la fonction objectif et 11 contraintes au niveau de la personne. Les contraintes au niveau de la personne sont les nombres totaux de personnes de 0 à 4 ans, de personnes de 5 à 17 ans, de personnes de 18 à 44 ans, de personnes de 45 à 64 ans, de personnes de 65 ans et plus, de personnes de sexe masculin, d'Hispaniques, de Noirs, d'Asiatiques et de personnes locales de leur logement.

Le tableau 4 contient les estimations-programmation quadratique généralisée pour le SP 1788, appelées

Tableau 4

Dénombrements au recensement, estimations ACE et estimations PQ avec contrôles ACE – DP 1788

Dénomb. au recensement	ACE	PQG	e.-l.(PQG)
Dénomb.	Dénomb.	Dénomb.	Dénomb.
(%)	(%)	(%)	(%)

Caractéristiques de l'unité de logement

Avec propres enfants	4 349	–	101,89	2,09
Sans propres enfants	3 685	–	101,66	3,07
1 à 4 personnes	6 785	–	102,03	2,03
5 personnes et plus	1 249	–	100,40	5,92
Unité occupée en location	2 559	–	104,57	2,62
Unité occupée en propriété	5 475	–	100,47	1,50
Total	8 034	–	101,78	1,22

Caractéristiques de la personne

0 à 4 ans	2 493	103,17	102,81	1,00
5 à 17 ans	6 339	103,09	103,08	0,96
18 à 44 ans	12 711	101,67	101,67	0,63
45 à 64 ans	3 028	100,26	100,33	0,59
65 ans et plus	574	99,48	98,95	0,70
Sexe masculin	12 473	102,18	102,01	0,68
Sexe féminin	12 672	101,74	101,82	0,62
Hispanique	2 385	104,95	104,91	1,09
Non hispanique	22 760	101,64	101,60	0,60
Noir	1 285	104,59	104,82	1,01
Blanc	22 372	101,69	101,69	0,61
Asiatique	257	100,00	101,95	1,95
Autres	1 231	104,47	102,92	1,14
Dans une unité en location	7 978	104,25	104,21	0,89
Dans une unité en propriété	17 167	100,89	100,84	0,68
Total	25 145	101,96	101,91	0,57

Bardsley et Chambers (1984) considèrent des estimateurs En utilisant $G(W)$ provenant de (20) et les 13 contrôles linéaires, nous obtenons les résultats présentés dans les deux dernières colonnes du tableau 2, sous l'entête « PQG ». Comme prévu les estimations s'approchent des totaux du recensement, parce que nous avons utilisé les totaux de marge du recensement comme contraintes. Les écarts relatifs en pourcentage entre l'estimation PQ et le dénombrement au recensement pour les 67 caractéristiques incluses dans $G(W)$ de (20) varient de -3,50 % à 3,75 %, avec environ 50 écarts inférieurs à 1 %.

Au tableau 3, nous comparons les poids d'échantillonnage obtenus par les deux méthodes de programmation à ceux produits par la méthode d'ajustement itératif proportionnel au niveau des ménages de l'USCB. Nous n'avons pas déterminé exactement le nombre et le type de contrôles utilisés pour l'ajustement itératif de l'USCB, parce qu'il dépend de l'exécution de la procédure de regroupement des cellules suivie par l'USCB et de certains fichiers de données provisoires qui n'est pas facile d'obtenir. Cependant, nous estimons que ce nombre est d'environ 67, parce que la procédure de regroupement utilise pour former les 67 cellules est essentiellement celle utilisée par l'USCB. La procédure PQ s'appuie sur celle utilisée par l'USCB. La procédure PQ s'appuie sur 82 contrôles et la procédure PQG, sur 90 contrôles. La fourchette de poids pour les deux méthodes de programmation quadratique est comparable, mais est plus étroite pour l'ajustement itératif proportionnel. Les écarts entre les trois sommes des carrés des poids sont modestes. Les valeurs de $g(W)$ sont également semblables, celle obtenue pour l'équation (20) étant la plus grande. La valeur de $g(W)$ est la quantité qui est minimisée par les poids figurant à la première ligne du tableau. La somme des carrés des poids pour la résolution PQ de l'équation (20) pourrait être réduite en diminuant la valeur de α_j dans la fonction objectif.

Tableau 3
Propriétés des poids des unités de logement de l'échantillon du questionnaire détaillé dans le SP 1788

Méthode	Poids minimal	Poids maximal	$\sum W_i^2$	$g(W)$
PQ avec $g(W)$ de (1)	1	26,5	78 028	326
72 contraintes				
PQ avec $G(W)$ de (20)	1	29,9	78 672	383
13 contraintes exactes				
Ajustement itératif proportionnel	4	22	77 000	369

Nous utilisons aussi les données provenant du Recensement de 1990 pour simuler la situation où les contrôles proviennent de dénombrements au recensement corrigés. Pour 1990, nous disposons des estimations au

Les 45 totaux de contrôle au niveau de la personne et les 22 totaux de contrôle au niveau de l'unité de logement obtenus en appliquant les règles de regroupement sont tels qu'une estimation de marge, comme le nombre total de personnes de sexe masculin, pourrait ne pas être contraainte de concorder avec le dénombrement. En outre, les procédures de regroupement de l'USCB produisent des contraintes au niveau de la personne et au niveau de l'unité de logement qui varient selon le secteur de pondération. Donc, nous avons considéré l'ajout de certains totaux de marge à l'ensemble de totaux de contrôle. Pour réduire l'effet de ces contrôles ajoutés sur les poids, nous avons remplacé les contraintes originales par des termes supplémentaires dans la fonction objectif. Ces termes sont les écarts entre les estimations finales et les totaux de contrôle. La fonction objectif devient

$$G(W) = g(W) + \sum_{j=1}^{67} \alpha_j \left(\sum_{i=1}^I W_i X_{ij} - X_j \right)^2, \quad (20)$$

où $g(W)$ est défini dans l'expression (1), les $\{X_j, j = 1, 2, \dots, 67\}$ sont l'ensemble de variables auxiliaires définissant les 45 contrôles au niveau de la personne et les 22 contrôles au niveau de l'unité de logement, et les α_j sont des constantes qu'il faut préciser. Le X_j pour la catégorie j du ménage i pour une caractéristique au niveau de la personne est le nombre de personnes appartenant à la catégorie j dans l'unité de logement. Le X_j pour une caractéristique au niveau de l'unité de logement vaut un si l'unité de logement a la caractéristique et est nul autrement. Dans notre application, la fonction est minimisée sous les contraintes de deux contrôles au niveau du ménage et de onze contrôles au niveau de la personne. Les contrôles au niveau de l'unité de logement sont les nombres d'unités occupées en location et d'unités occupées en propriété. Les contrôles au niveau de la personne sont les nombres de personnes de 0 à 4 ans, de personnes de 5 à 17 ans, de personnes de 18 à 44 ans, de personnes de 45 à 64 ans, de personnes de 65 ans et plus, de personnes de sexe masculin, de Noirs, de Blancs, d'Asiatiques, d'Hispaniques et de locales. Les constantes α_j sont $10[W_{(2)}^{-1} - \sigma_j^2]^{-1}$, où $W_{(2)} = 8,95$ est la moyenne des $W_{(2)}'$, $\sigma_j^2 = P_j(1 - P_j)$ et P_j est la proportion de la population comprise dans la cellule j . Les α_j minimisent l'erreur quadratique moyenne d'un total estimé s'il n'y avait qu'une seule variable de contrôle et que le carré de la corrélation entre la variable de contrôle et la variable dépendante était d'environ 0,9. Donc, la pression exercée par la fonction afin que l'estimation finale s'approche du total de contrôle est considérable.

La résolution de (20) par programmation quadratique donne un type d'estimateur par régression. Voir Fuller (2002) et Fuller et Isaki (2001). Rao et Singh (1997) et

dénombrément et l'estimation PQ pour les caractéristiques du ménage est comparable à celle observée pour les estimations de l'USCB fondées sur les ménages et supérieure à celle observée pour les estimations de l'USCB fondées sur les dénombrements de personnes. Pour les dénombrements de personnes, les estimations PQ s'approchent généralement plus des dénombrements au recensement que l'une ou l'autre estimation par la méthode itérative du quotient de l'USCB.

L'écart le plus important entre l'estimation PQ et le dénombrement au recensement relativement à l'erreur-type est celui observé pour l'estimation du nombre de ménages avec présence des propres enfants, où la différence est d'environ 1,6 erreur-type. La majorité des estimations PQ rapport aux valeurs de contrôle.

Tableau 2
Caractéristiques estimées des unités de logement occupées et des personnes pour le SP 1788

Caractéristiques de l'unité de logement	Dénombrement au recensement					
	$\frac{\text{Est.}(L^*)}{\text{Dénomb.}} (\%)$	$\frac{\text{Est.}(P^{**})}{\text{Dénomb.}} (\%)$	$\frac{PQ^{\dagger}}{\text{Dénomb.}} (\%)$	$\frac{e.-l.(PQ)}{\text{Dénomb.}} (\%)$	$\frac{PQ^{\ddagger}}{\text{Dénomb.}} (\%)$	$\frac{e.-l.(PQG)}{\text{Dénomb.}} (\%)$
Avec propres enfants	4 349	100,18	100,45	100,21	100,18	0,14
Sans propres enfants	3 685	99,78	99,67	99,76	99,78	0,16
1 à 4 personnes	6 785	100,00	100,57	100,04	100,07	0,05
5 personnes et plus	1 249	100,00	97,51	99,76	99,60	0,30
Unité occupée en location	2 559	100,00	95,97	100,00	99,92	0,16
Unité occupée en propriété	5 475	100,00	102,02	100,00	100,04	0,08
Caractéristiques de la personne						
0 à 4 ans	2 493	101,92	97,95	98,84	99,96	0,29
5 à 17 ans	6 339	103,91	101,07	100,63	99,98	0,18
18 à 44 ans	12 711	99,50	99,69	100,01	100,00	0,06
45 à 64 ans	3 028	101,65	101,95	99,90	99,97	0,09
65 ans et plus	574	81,18	93,73	100,17	100,00	0,27
Sexe masculin	12 473	99,95	99,64	100,06	99,98	0,09
Sexe féminin	12 672	101,43	100,36	99,95	100,01	0,09
Hispanique	2 385	95,38	103,40	99,96	99,87	0,38
Non hispanique	22 760	101,25	99,64	100,03	100,00	0,10
Noir	1 285	101,08	101,79	100,86	99,77	0,54
Blanc	22 372	100,69	99,91	100,03	100,00	0,10
Asiatique	257	92,60	80,05	96,83	99,76	0,50
Autre	1 231	101,94	103,89	105,84	100,78	1,75
Dans une unité en location	7 978	102,04	95,41	100,01	99,92	0,19
Dans une unité en propriété	17 167	100,06	102,13	100,00	100,02	0,13

* Poids de l'USCB pour les ménages
** Poids PQ avec 82 contraintes
† PQ généralisée avec 13 contraintes et la fonction objectif (20)

Maintenant, supposons que $E\{\hat{V}_{AA}\} = V_{AA}$, $E\{\hat{\beta}^{(i)}\} = \beta$ et que \hat{V}_{AA} est indépendant de $\hat{\beta}^{(i)}$,

$$E\{\hat{\beta}^{(i)}\hat{V}_{AA}\hat{\beta}^{(i)}\} = \beta'V_{AA}\beta + \text{tr}\{\hat{\beta}^{(i)}\}V_{AA}\hat{\beta}^{(i)};$$

où $\text{tr}\{V_{AA}\}$ est la trace de la matrice. Il s'ensuit que

$$E\left\{r^{-1}(r-1)\sum_{i=1}^l(\hat{\theta}^{R(i)}-\hat{\theta}^R)^2\right\} = E\left\{r^{-1}(r-1)\sum_{i=1}^l(\hat{\theta}^{R(i)}-\hat{\theta}^R)^2\right\} + \beta'V_{AA}\beta + O(n^{-2}), \quad (19)$$

où nous supposons que $\text{tr}\{V_{AA}\} = O(n^{-1})$ et $\text{tr}\{V^{(i)}\hat{\beta}^{(i)}\} = O(n^{-1})$, où n est la taille d'échantillon. Le premier terme du deuxième membre de l'égalité (19) est l'espérance de l'estimateur de la variance pour la variance due à l'échantillonnage pour le questionnaire détaillé du recensement. Le deuxième terme est la contribution de la variance à l'erreur dans les estimations d'après l'ACE de la variance totale. Donc, l'estimateur de la variance fondé sur $\hat{\theta}^{R(i)}$ estime les deux composantes de la variation. Notons que la matrice des covariances estimée pour les contrôles est \hat{V}_{AA} , comme elle devrait l'être.

4. RÉSULTATS NUMÉRIQUES

Nous utilisons le fichier de données du Recensement de 1990 de l'USCB pour illustrer l'application de la méthode de programmation quadratique à des données réelles. Le fichier contient des données sur les ménages et sur les personnes dans les ménages, ainsi que les poids applicables aux données du questionnaire détaillé tels qu'élaborés pour le Recensement des États-Unis de 1990. Donc, le fichier fournit des données appropriées pour comparer les propriétés de la méthode de pondération des données du questionnaire détaillé de 1990 de l'USCB à celles de la méthode de programmation quadratique.

L'USCB réalise la pondération de l'échantillon du questionnaire détaillé selon le secteur de pondération, ce dernier contenant habituellement de deux à trois mille unités de logement. En 1990, aux États-Unis, il existait environ 56 000 secteurs de pondération. Pour notre travail numérique, nous avons choisi le secteur de pondération (SP) 1788 qui contient 8 034 unités de logement occupées et 25 145 personnes.

Nous donnons au tableau 2 les estimations pour certaines caractéristiques aux niveaux de la personne et de l'unité de logement pour le secteur de pondération 1788. Les caractéristiques figurant dans le tableau, sauf le nombre d'unités louées, ont été suggérées par les spécialistes du domaine à l'USCB. Dans le tableau 2, Est.(L) est l'estimation pondérée d'après l'échantillon du questionnaire détaillé calculée en utilisant les poids des logements, Est.(P) est l'estimation pondérée d'après l'échantillon du questionnaire détaillé calculée en utilisant les poids des personnes qui sont présentes au tableau 2 ont été créées en utilisant le poids du chef de ménage comme poids de l'unité de logement. Chaque unité occupée ne contient qu'un seul chef de ménage. La méthode axée sur le chef de ménage est appelée *méthode de la personne principale* par Alexander (1987). Toutes les estimations qui figurent dans le tableau sont exprimées en pourcentage du dénombrement au recensement.

Les estimations calculées par les deux méthodes de l'USCB peuvent différer de plusieurs points de pourcentage, les écarts entre Est.(P) et Est.(L) étant appréciables pour les unités louées, les personnes de 0 à 4 ans, les personnes de 65 ans et plus, les Hispaniques, les Asiatiques et les personnes locales. L'estimation Est.(L) pour les personnes locales s'approche plus de 100 que l'estimation Est.(P).

Les règles de regroupement des cellules produisent 45 contrôles au niveau de la personne et 22 contrôles au niveau de l'unité de logement pour le SP 1788. Un exemple de contrôle au niveau de la personne est le nombre total d'hommes noirs non hispaniques de 65 ans et plus, tandis qu'un exemple de contrôle au niveau de l'unité de logement est le nombre total d'unités de logement dont le propriétaire est un blanc non hispanique. Le nombre total de personnes noires est un contrôle implicite pour le SP 1788. Nous avons ajouté des contrôles pour les nombres totaux de personnes de 18 à 44 ans, de personnes de 45 à 64 ans, de personnes de sexe masculin, de locataires et d'unités de logement occupées en location pour la programmation quadratique. À part les contrôles susmentionnés, aucune autre caractéristique figurant au tableau 2 n'est utilisée comme contrôle dans cette méthode. Les estimations PQ et les erreurs-types des estimations PQ sont présentées, exprimées en pourcentage du dénombrement au recensement.

Les estimations PQ et les erreurs-types des estimations PQ sont présentées, exprimées en pourcentage du dénombrement au recensement, dans les quatrième et cinquième colonnes du tableau 2. La concordance entre le

$$V\{\theta^w\} = (r-1)^{-1}(r-2)V_R\{\theta_R\} + r^{-1}\sum_{j=1}^r(\theta^{w(j)} - \bar{\theta}^w)^2. \quad (11)$$

3.2 Répétitions pour les contrôles d'après l'ACE

Nous avons modifié les répétitions pour les estimations produites d'après des valeurs de contrôle provenant de l'ACE de sorte que les variances estimées contiennent une composante pour l'erreur contenue dans les estimations ACE. Nous avons assigné les données comprises dans un secteur de pondération à 67 répétitions, où 67 est le nombre de contrôles. La procédure exige que le nombre de répétitions soit égal ou supérieur au nombre de contrôles s'il faut que la matrice des covariances des totaux de contrôle soit reproduite. Il est possible d'utiliser un plus grand nombre de répétitions que de contrôles. Voir Fuller (1998).

L'estimateur du total d'une caractéristique pour le questionnaire détaillé est une forme d'estimateur par régression utilisant les chiffres de l'ACE comme contrôles. Nous écrivons l'estimateur du total fondé sur les valeurs réelles des poids sous la forme

$$\hat{\theta}_R = \hat{X}_A \hat{\beta}, \quad (12)$$

où \hat{X}_A est le vecteur d'estimation ACE et $\hat{\beta}$ est le coefficient de régression calculé d'après les données provenant du questionnaire détaillé.

Soit \hat{V}_{AA} la matrice des covariances de dimensions $r \times r$ du vecteur de contrôles ACE, où \hat{V}_{AA} est estimée en tant qu'élément du processus de l'ACE, et $r = 67$. Soit $\lambda_1, \lambda_2, \dots, \lambda_r$ les racines de \hat{V}_{AA} et soit

$$\hat{Q} \hat{V}_{AA} \hat{Q} = \Lambda, \quad (13)$$

où $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$, et \hat{Q} est la matrice composée des vecteurs de caractéristiques de \hat{V}_{AA} .

Rappelons que

$$\hat{V}_{AA} = \hat{Q} \Lambda \hat{Q}'$$

et

$$\hat{V}_{AA} = \sum_{j=1}^r \mathbf{q}_{\cdot j} \lambda_j \mathbf{q}_{\cdot j}' = \sum_{j=1}^r \mathbf{z}_{\cdot j} \mathbf{z}_{\cdot j}' \quad (14)$$

où $\mathbf{q}_{\cdot j}$ est la j^{e} colonne de \hat{Q} et $\mathbf{z}_{\cdot j} = \lambda_j^{1/2} \mathbf{q}_{\cdot j}$.

En utilisant le résultat (14), nous construisons les contrôles pour les r répétitions tels que

$$\hat{X}_{A(i)} = \hat{X}_A + c \mathbf{z}_{\cdot i}, \quad i = 1, 2, \dots, r, \quad (15)$$

où \hat{X}_A est le vecteur de ligne des valeurs de contrôles originels et c est une constante. La constante c est déterminée de telle façon que l'espérance de la somme des carrés des écarts du jackknife pour les éléments du vecteur \hat{X} soient les éléments de la diagonale de \hat{V}_{AA} . Dans notre application, la constante c est $(r-1)^{-1/2} r^{1/2}$ et

$$(r-1)^{-1} \sum_{j=1}^r c^2 \mathbf{z}_{\cdot j} \mathbf{z}_{\cdot j}' \quad (16)$$

Donc, si la caractéristique « estimée » correspond à l'un des contrôles utilisés dans la PQ, la méthode du jackknife donne la variance estimée d'après l'ACE pour cette caractéristique. Les $\mathbf{z}_{\cdot j}$ sont assignés au hasard aux r répétitions.

En utilisant la représentation de la régression, nous écrivons l'estimateur pour la i^{e} répétition sous la forme

$$\hat{\theta}^{R(i)} = \hat{X}^{A(i)} \hat{\beta}^{(i)}$$

$$= \hat{X}_A \hat{\beta}^{(i)} + (\hat{X}^{A(i)} - \hat{X}_A) \hat{\beta}^{(i)}$$

$$=: \hat{\theta}^{R(i)} + c \mathbf{z}_{\cdot i}' \hat{\beta}^{(i)}, \quad (17)$$

où $\hat{\theta}^{R(i)}$ est l'estimateur avec valeurs réelles des poids calculé avec le i^{e} groupe supprimé en utilisant $\hat{X}^{A(i)}$ comme vecteur de contrôles, $\hat{\beta}^{(i)}$ est le coefficient de régression calculé avec le i^{e} groupe supprimé, et $\hat{\theta}^{R(i)}$ est l'estimateur avec valeurs réelles des poids calculés avec le i^{e} groupe supprimé en utilisant \hat{X}_A comme vecteur de contrôles.

Alors

$$\hat{\theta}^{R(i)} - \bar{\theta}_R = \hat{\theta}^{R(i)} - \bar{\theta}_R + c \mathbf{z}_{\cdot i}' \hat{\beta}^{(i)}.$$

Puisque les $\mathbf{q}_{\cdot j}$ sont assignées aux répétitions au hasard, l'espérance de l'estimateur de la variance de la répétition pour l'estimateur avec valeurs réelles des poids fondés sur les contrôles ACE est

$$E\left\{V_R(\bar{\theta}^{R(i)})\right\} = E\left\{r^{-1} \sum_{i=1}^r (\theta^{R(i)} - \bar{\theta}^R)^2\right\}$$

$$= E\left\{r^{-1} \sum_{i=1}^r (\theta^{R(i)} - \bar{\theta}^R)^2\right\} + E\left\{\hat{\beta}^{(i)} \hat{V}_{AA} \hat{\beta}^{(i)}\right\}. \quad (18)$$

3. ESTIMATION DE LA VARIANCE

Nous supposons que l'erreur résultant de l'arrondissement est indépendante du choix du groupe éliminé, hypothèse raisonnable, étant donné que la suppression produit un ensemble entièrement nouveau de poids qui doivent être arrondis. Alors

$$E\{(\bar{\theta}^{w(t)} - \bar{\theta}^{R(t)})^2\} = E\{(\bar{\theta}^{R(t)} - \bar{\theta}^{w(t)})^2\} + E\{(\bar{\theta}^{w(t)} - \bar{\theta}^{R(t)})^2\} \quad (7)$$

Supposons que la moyenne de $\bar{\theta}^{R(t)}$ est égale à $\bar{\theta}^R$. Alors, le dernier terme de (7) est un écart de répétition pour la différence entre les estimations réelles et arrondies. Alors

$$E\left\{\left[\left(\bar{\theta}^{w(t)} - \bar{\theta}^{R(t)}\right) - \left(\bar{\theta}^w - \bar{\theta}^R\right)\right]^2\right\} = V\{(\bar{\theta}^{w(t)} - \bar{\theta}^w) + V\{(\bar{\theta}^{R(t)} - \bar{\theta}^R)\} \quad (8)$$

où $V\{\bar{\theta}^{w(t)} - \bar{\theta}^{R(t)}\}$ est la variance due à l'arrondissement pour un échantillon de $r-1$ groupes et $V\{\bar{\theta}^w - \bar{\theta}^R\}$ est la variance due à l'arrondissement pour un échantillon de r groupes. En obtenant (8), nous supposons que la variance est égale à la variance pour $r-1$ groupes multipliée par $r^{-1}(r-1)$. Donc

$$E\left\{(\bar{\theta}^{w(t)} - \bar{\theta}^{R(t)})^2\right\} = E\left\{(\bar{\theta}^{w(t)} - \bar{\theta}^w)^2\right\} + V\{(\bar{\theta}^{R(t)} - \bar{\theta}^R)\}$$

$$+ V\{\bar{\theta}^w - \bar{\theta}^R\}, \quad (9)$$

où

$$V\{\bar{\theta}^R\} = (r-1) \sum_{i=1}^r (\bar{\theta}^{R(i)} - \bar{\theta}^R)^2$$

est l'estimateur par le jackknife de la variance pour l'estimateur obtenu d'après les poids réels. Alors, un estimateur de la variance due à l'arrondissement est

$$V\{\bar{\theta}^w - \bar{\theta}^R\} =$$

$$\left[(r-1) \sum_{i=1}^r (\bar{\theta}^{w(i)} - \bar{\theta}^w)^2 - r(r-1)^{-1} V_R\{\bar{\theta}^R\} \right] \quad (10)$$

D'après ces résultats, la variance estimée pour l'estimateur arrondi est

Nous avons estimé la variance des estimations calculées d'après les données du questionnaire détaillé par la méthode du jackknife. Pour les résultats numériques obtenus en prenant des valeurs de contrôle provenant du recensement, nous avons formé 16 échantillons répétés. Nous avons choisi ce nombre pour des raisons de commodité, mais nous aurons pu en utiliser un plus grand. Nous avons tiré l'échantillon du questionnaire détaillé selon le numéro d'identification du recensement à l'intérieur des îlots et nous avons formé 16 échantillons répétés correspondant aux 16 échantillons systématiques prélevés à intervalle de un sur seize. Pour les estimations fondées sur des valeurs de contrôle provenant de l'ACE, nous avons formé 67 échantillons répétés.

3.1 Échantillons répétés pour les contrôles d'après le recensement

Les répliques jackknife sont créées en supprimant le i^e groupe d'éléments, en calculant les poids par programmation quadratique et en arrondissant ceux-ci à des nombres entiers. À cause de l'arrondissement, il est nécessaire de modifier la méthode habituelle d'estimation de la variance par le jackknife. Pour isoler l'effet de l'arrondissement, nous considérons l'estimation répétée établie d'après les valeurs réelles des poids. Soit

$\bar{\theta}^w$ = l'estimateur d'échantillon avec poids arrondis à des nombres entiers,
 $\bar{\theta}^R$ = l'estimateur d'échantillon avec valeurs réelles des poids,
 $\bar{\theta}^{R(i)}$ = l'estimation répétée jackknife avec suppression du i^e groupe et valeurs réelles des poids,
 $\bar{\theta}^{w(i)}$ = l'estimation répétée jackknife avec suppression du i^e groupe et poids arrondis à des entiers,

et soit

$$\bar{\theta}^w = r^{-1} \sum_{i=1}^r \bar{\theta}^{w(i)}, \quad (5)$$

où r est le nombre total de répétitions. Alors, l'écart du jackknife pour l'estimateur avec poids entiers peut être décomposé comme suit

$$\bar{\theta}^{w(t)} - \bar{\theta}^w = \bar{\theta}^{R(t)} - \bar{\theta}^R + \left[\bar{\theta}^{w(t)} - \bar{\theta}^w - (\bar{\theta}^{R(t)} - \bar{\theta}^R) \right]. \quad (6)$$

avons fixé la valeur de K à 48, mais cette borne n'a jamais été atteinte. Par contre la borne inférieure de un l'a été. Nous avons utilisé la sous-routine FORTRAN provenant de IMSL pour résoudre le problème de programmation quadratique. D'autres programmes, tels que LCP ou SAS®/IML, peuvent aussi être utilisés.

La méthode de pondération des données du questionnaire détaillé utilisée à l'heure actuelle par l'USCB consiste à utiliser la méthode itérative du quotient (raking ratio) des dénombrements obtenus d'après l'échantillon du questionnaire détaillé initialement pondérés d'après les dénombrements du recensement pour les catégories de contrôle. La pondération est réalisée pour des subdivisions du pays, appelées secteurs de pondération, séparément pour les caractéristiques des personnes et celles des ménages. Pour le questionnaire détaillé, les taux nominaux d'échantillonnage sont sur deux, un sur six et un sur huit. Les poids nominaux d'échantillonnage, sont représentés par $W'_{(1)}$. Un deuxième ensemble de poids, représentés par $W'_{(2)}$, correspond aux taux d'échantillonnage réalisés, calculés pour les cellules, où les cellules doivent contenir au moins cinq ménages échantillonnés. Pour des précisions sur les procédures suivies par l'USCB, voir Schindler et coll. (1992).

Puisque nous avons l'intention de comparer la méthode itérative du quotient et la méthode de programmation quadratique, nous utilisons la plupart des catégories établies par l'USCB au niveau des personnes et des ménages comme totaux de contrôle X_j et Z_j dans la programmation quadratique, mais en apportant certains changements. Par exemple, nous retenons toutes les catégories individuelles âge-race-sexe, mais nous n'utilisons pas celle fondée sur les taux d'échantillonnage nominaux.

Nous utilisons les spécification de l'USCB pour déterminer si une catégorie de cellules doit être retenue en tant que contrôle distinct ou être combinée à d'autres cellules, et nous utilisons la méthode suivie par l'USCB pour déterminer quelles cellules doivent être combinées. Cette approche nous permet de tirer parti de l'expérience de l'USCB et de réduire au minimum les écarts entre l'ensemble de totaux de contrôle appliqués par l'USCB aux données du questionnaire détaillé et l'ensemble utilisé pour la méthode de programmation quadratique. La procédure utilisée pour définir $W'_{(2)}$ est décrite en annexe.

Deux ensembles de totaux de contrôle peuvent être utilisés pour construire les poids à utiliser pour le questionnaire détaillé du Recensement des États-Unis de 2000. Le premier de ces ensembles est celui des valeurs de contrôle obtenues d'après le questionnaire abrégé du Recensement de 2000. Autrement dit, les totaux de contrôle indépendants qu'il convient de maintenir lors de la

pondération des données du questionnaire détaillé sont ceux calculés d'après les données du questionnaire abrégé du recensement. Lorsqu'on utilise les données de recensement comme contrôle, les catégories de contrôle au niveau de la personne (X_j) incluent une classification croisée de l'âge et du sexe-race/ethnicité. D'autres caractéristiques, comme le mode d'occupation du logement, sont utilisées à titre de contrôles supplémentaires. La majorité des catégories de contrôle au niveau du ménage (Z_j) sont définies par classification croisée du type de ménage (par exemple famille avec enfants de moins de 18 ans) et de la taille du ménage (par exemple nombre de personnes dans la famille). Les variables Z_j incluent aussi la race/ethnicité du chef de ménage recoupée par le mode d'occupation du logement. L'autre ensemble de totaux de contrôle que nous pouvons utiliser pour le Recensement de 2000 est l'enquête d'estimations obtenues d'après l'enquête postcensitaire, appelée Enquête d'évaluation de l'exactitude et de la couverture (ACE pour *Accuracy and Coverage Evaluation*). L'enquête ACE est conçue pour estimer uniquement les caractéristiques au niveau individuel et les X_j incluent les contrôles âge-sexe-race/ethnicité-mode d'occupation du logement.

La dernière étape de la pondération des données du questionnaire détaillé est l'arrondissement des poids obtenus W_i en nombres entiers. L'utilisation de poids entiers évite les écarts entre les ensembles d'estimations dus à l'arrondissement des estimations calculées d'après les valeurs réelles. Nous avons regroupé les unités de logement échantillonnées selon la race/ethnicité du chef de ménage et selon le mode d'occupation du logement. Puis, dans chaque groupe, nous avons tiré l'échantillon selon le type de famille et selon la taille du ménage. Ensuite, nous avons arrondi les poids à un nombre entier en utilisant la méthode « du cumuli et de l'arrondissement » qui est illustrée au tableau I. Elle comprend le calcul des sommes partielles (cumuli) des poids, tel qu'indiqué dans la colonne dominant les poids cumulés (PC), puis l'arrondissement des sommes partielles tel qu'indiqué dans la colonne PCA. Le poids entier pour l'élément i est égal à la différence entre les entées successives $i - 1$ et i dans la colonne PCA.

Tableau I

Illustration de la méthode du cumuli et de l'arrondissement

Unité	Poids initial	PC	PCA	Poids entier
1	3,333	3	3	3
2	2,500	5,833	6	3
3	1,428	7,261	7	1
4	1,250	8,511	9	2
5	1,111	9,622	10	1
6	5,021	14,643	15	5

contrôle particuliers et (iv) produisent un estimateur convergent par rapport au plan de sondage. À part l'application de bornes, les poids obtenus par programmation quadratique sont ceux d'un estimateur par régression simple. Nous commençons par décrire la forme mathématique de la PQ, puis nous discutons de l'exécution. Soit

(i) $\{W_i; i = 1, 2, \dots, n\}$ l'ensemble des poids initiaux des unités de l'échantillon ou i dénote le i^{e} ménage de cet échantillon;

(ii) $\{W_i^{(2)}; i = 1, 2, \dots, n\}$ l'ensemble de poids initiaux des unités de logement;

$$(m) \quad X_{ij}, j = 1, 2, \dots, m_d, i = 1, 2, \dots, n \text{ observation sur}$$

la j variable de contrôle au niveau de la personne pour le 1^{er} ménage de l'échantillon.

‘*HOHHEIM*’ I CAN RECOGNISE. I CAN

$$Z^{f_i}, f_i = 1, 2, \dots, m_i, i = 1, 2, \dots, n \text{ observation sur}$$

al pond egiwam na naan na eionwot ap eibabw / a

‘**ПОПОВИЧУКА И ДОЗВРНАЦИ**’

(v) $X_j, j = 1, 2, \dots, m_d$ la j^{e} variable de contrôle au niveau

de la personne;

vi) Z_j , $j = 1, 2, \dots, m_h$ la j^{e} variable de contrôle au niveau

du ménage.

$$x = 0.40, 0.30, 0.20, 0.10, 0.05, 0.02, 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.00001$$

La méthode de programmation quadratique consiste à

objectif quadratique soumis à des contraintes linéaires.

Dans notre application, nous minimisons

$$(I) \quad {}_{-1}\left[\begin{smallmatrix} i \\ (2) \end{smallmatrix} M \right]_2 \left(\begin{smallmatrix} i \\ (2) \end{smallmatrix} M - \begin{smallmatrix} i \\ (2) \end{smallmatrix} M \right) \sum_n = (M)^{\otimes}$$

sous les contraintes

$$\sum_n W_n X_n = X, \quad \text{pour } i = 1, 2, \dots, m, \quad (2)$$
$$d = \frac{1}{2} \left(\frac{1}{f} + \frac{1}{f'} \right) \quad \text{if } f \neq f' \quad \text{if } f = f' \quad d = f$$
$$\frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i} = 1$$

(c) $\langle u, \dots, z, 1 \rangle = f \bmod \langle z = 1, z^2, \dots, z^l \rangle$

$$(4) \quad Y \leq M \leq I$$

où les sommations sont faites sur l'ensemble des unités de

logement comprises dans l'échantillon répondant au

questionnaire détaillé. Notons que nous imposons une borne

inférieure égale à l'unité aux poids appliqués aux ménages

répondant au questionnaire détaillé. Cette contrainte signifie

que tout élément de l'échantillon devrait au moins être

QUADRATIQUE

$$(1) \quad {}_{-1}^{-1}[^i\mathcal{M}]_z \left({}_{(z)}^i\mathcal{M} - {}^i\mathcal{M} \right) \sum_{u=1}^l (\mathcal{M})^{\delta}$$

(iv) $Z_{ji}^f, j = 1, 2, \dots, m, i = 1, 2, \dots, n$ l'observation sur la j^{e} variable de contrôle au niveau du ménage pour le i^{e} ménage et l'échantillon;

(v) $X_j^f, j = 1, 2, \dots, m_p$ la j^{e} variable de contrôle au niveau de la personne;

(vi) $Z_j^f, j = 1, 2, \dots, m_h$ la j^{e} variable de contrôle au niveau du ménage.

La méthode de programmation quadratique consiste à rechercher les $W_i^f, i = 1, 2, \dots, n$ qui minimisent une fonction objectif quadratique soumise à des contraintes linéaires, dans notre application nous minimisons

(i) $\{W_i, i = 1, 2, \dots, n\}$ l'ensemble des poids finaux des unités de logement, où i dénote le i^{e} ménage de l'échantillon du questionnaire détaillé et n est la taille de cet échantillon;
 (ii) $\{W_{(2)}^i, i = 1, 2, \dots, n\}$ l'ensemble de poids initiaux des unités de logement;
 (iii) $X^{j,i}, j = 1, 2, \dots, m^d, i = 1, 2, \dots, n$ l'observation sur

particuliers et (1) produisent un estimateur par programmation de bornes, les poids obtenus par régression simple. Nous commençons par décrire la forme mathéma-

Pondération de données d'échantillon reposant sur des contrôles indépendants

CARY T. ISAKI, JULIE H. TSAI et WAYNE A. FULLER¹

RÉSUMÉ

Lors du Recensement de la population et du logement des États-Unis, un échantillon d'environ un sixième des ménages reçoit une version longue du questionnaire de recensement appelée questionnaire détaillé. Les autres ménages reçoivent une version courte appelée questionnaire abrégé. Nous recourons à l'ajustement itératif proportionnel, au moyen de certains totaux de contrôle provenant du questionnaire abrégé, pour créer deux ensembles de poids pour l'estimation d'après les données provenant du questionnaire détaillé. L'un pour les personnes et l'autre pour les ménages. Nous décrivons une méthode de calcul des poids fondée sur la programmation quadratique qui produit une pondération des ménages telle que la somme pondérée des caractéristiques individuelles et celle des caractéristiques des ménages concordent étroitement avec les totaux de contrôle fondés sur le questionnaire abrégé. La méthode s'applique de façon générale aux situations où la pondération doit être établie de façon à satisfaire aux bornes de taille, ainsi qu'aux contraintes de concordance des sommes avec les totaux de contrôle. Nous décrivons l'application à la situation où les totaux de contrôle sont des estimations avec une matrice des covariances estimée.

MOTS CLÉS : Méthode itérative du quotient (raking ratio); régression; programmation quadratique; rajustement de la couverture; poids entiers; secteur de pondération.

1. INTRODUCTION

À condition de connaître les totaux pour certaines caractéristiques, il est courant que les statisticiens d'enquête utilisent cette information dans les estimateurs par stratification a posteriori, par le quotient ou par la régression. Les totaux connus de caractéristiques sont parfois appelés contrôles indépendants, parce qu'ils sont calculés dans un autre contexte que celui de l'enquête. L'utilisation de contrôles indépendants a tendance à réduire la variance de la plupart des estimations. En outre, ces contrôles permettent souvent de pallier les problèmes de couverture dans les enquêtes. Voir Deville et Särndal (1992) et Fuller (2002).

Le recensement décennal des États-Unis s'appuie sur un échantillon pour la mesure de certaines caractéristiques. Le questionnaire portant sur ces caractéristiques est appelé questionnaire détaillé et l'échantillon auquel il s'applique est un échantillon aléatoire d'adresses. Le questionnaire comprend les questions qui sont posées à tous les individus (information du questionnaire abrégé), ainsi que des questions sur un ensemble de caractéristiques supplémentaires. Lors des recensements antérieurs, on a utilisé la méthode itérative du quotient (raking ratio) d'après des contrôles fondés sur l'information du questionnaire abrégé pour produire les poids à appliquer à l'échantillon répondant au questionnaire détaillé. Deux ensembles de poids ont été créés ainsi, l'un pour les caractéristiques des personnes et l'autre pour celles des unités de logement.

L'ensemble de catégories utilisées pour la pondération au niveau des personnes était une classification des individus selon la race, l'origine hispanique, l'âge et le sexe, le type de famille et la taille du ménage. Pour les ménages, les catégories étaient la classification croisée de la race selon l'origine hispanique du chef de ménage selon le mode d'occupation du logement selon le type et la taille du logement. Lors de la pondération des données provenant du questionnaire détaillé du Recensement de 1990, les pertotales indépendantes à réduire la variance de la plupart des estimations. En outre, ces contrôles permettent souvent de pallier les problèmes de couverture dans les enquêtes. Voir Deville et Särndal (1992) et Fuller (2002).

Le recensement décennal des États-Unis s'appuie sur un échantillon pour la mesure de certaines caractéristiques. Le questionnaire portant sur ces caractéristiques est appelé questionnaire détaillé et l'échantillon auquel il s'applique est un échantillon aléatoire d'adresses. Le questionnaire comprend les questions qui sont posées à tous les individus (information du questionnaire abrégé), ainsi que des questions sur un ensemble de caractéristiques supplémentaires. Lors des recensements antérieurs, on a utilisé la méthode itérative du quotient (raking ratio) d'après des contrôles fondés sur l'information du questionnaire abrégé pour produire les poids à appliquer à l'échantillon répondant au questionnaire détaillé. Deux ensembles de poids ont été créés ainsi, l'un pour les caractéristiques des personnes et l'autre pour celles des unités de logement.

La pondération des données provenant du questionnaire détaillé en utilisant l'information du questionnaire abrégé de recensement fait partie intégrante du Recensement de la population et du logement du Canada. Contrairement à la procédure utilisée par le U.S. Census Bureau (USCB), celle utilisée par Statistique Canada consiste à produire un ensemble unique de poids des ménages en s'appuyant sur l'estimation par régression. Voir Bankier, Houle et Luc (1997). Si les poids initiaux obtenus par la méthode de

1 Cary T. Isaki et Julie H. Tsai, U.S. Bureau of the Census, Statistical Research Division, Washington, D.C. 20233, États-Unis
Courtnei : Julie Hsu-Ling T. Tsai@census.gov; Wayne A. Fuller, Iowa State University, Department of Statistics, 221 Smedecor Hall, Ames, Iowa 50011, États-Unis.

La modification proposée au point iii) pour traiter la non-réponse totale est très importante et elle élargit le champ d'application de notre méthode. Comme l'ont fait remarquer Skinner, Kott et Fuller (2002), il est important de retenir les poids g dans l'estimation de la variable dans les situations où les valeurs limites des estimateurs X diffèrent des totaux de contrôle correspondants X , comme dans le cas de la non-réponse ou de la non-couverture. Notre méthode tient compte automatiquement des poids g et peut produire des estimateurs de la variance convergents dans de tels cas. Les résultats empiriques obtenus par Skinner avec d'Avrigo dans ce contexte sont fort intéressants. Le cas des estimateurs de la variance pour d'autres estimateurs par calage étudiés à la section 3.4, comparativement aux estimateurs habituels de la variance qui remplaçaient $d_k(s)f(x_k^T \lambda)$ dans l'expression de B_k par $d_k(s)$ ou par $w_k(s)$, doit être étudié plus en profondeur, comme le souligne Skinner.

Il convient de noter que la non-réponse totale est habituellement traitée comme un échantillonnage de deuxième phase (par exemple, échantillonnage de Poisson avec probabilités de réponse inconnues) et Skinner remarque que notre méthode peut donner lieu à des estimateurs de la variance convergents, même quand les estimateurs sont fondés uniquement sur les probabilités d'inclusion sans l'échantillon. Cependant, les totaux de contrôle X sont nécessaires pour obtenir des estimateurs valides du total Y , sous certaines hypothèses concernant les probabilités de réponse (Fuller 2002, équation (8.4)). Nous avons énoncé notre méthode au traitement de la non-réponse totale pour tenir compte de la non-réponse totale et de l'imputation pour tenir compte de la non-réponse partielle lorsqu'on ne dispose pas de totaux de contrôle, en supposant que la réponse est uniforme à l'intérieur des classes (Demnati et Rao 2002). Les estimateurs de la variance résultants sont naturellement plus complexes que la modification proposée par Skinner pour la non-réponse totale en présence de totaux de contrôle.

BIBLIOGRAPHIE

En ce qui concerne le point iv) sur l'utilisation éventuelle de la différenciation numérique pour calculer les variables linéarisées z_k^p , Woodruff et Cauley (1976) ont utilisé une méthode de ce genre pour calculer les dérivées $\partial g(a)/\partial a_i|_{a=g}$ données dans (1.4) quand $\theta = g(X)$. Skinner propose de perturber chaque poids $d_k(s)$ à tour de rôle, puis de recalculer θ , par exemple, en le remplaçant par une quantité fixe θ plus petite que la valeur minimale de $d_k(s)$, $k \in s$. Il présume que l'approche proposée produira des estimateurs de la variance fort semblables à ceux obtenus par différenciation analytique. Il serait utile d'étudier les propriétés statistiques de l'approche proposée de la différenciation analytique de $f(d(s))$ par rapport aux poids $d_k(s)$.

Nous espérons que les discussions de Kott, Shah et Skinner susciteront d'autres travaux sur l'approche de l'estimation de la variance présentée dans notre article.

BERGGER, Y.G., et SKINNER, C.J. (2003). Variance estimation for a low income proportion. *Applied Statistics*, 52, 457-468.

DEMNATI, A., et RAO, J.N.K. (2002). Linearization variance estimators for survey data with missing responses. *Proceeding of the Section Survey Research Methods*, American Statistical Association, 736-740.

FRANCISCO, C.A., et FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

SHAO, J. (1991). L-statistics in complex problems. Rapport technique, University of Ottawa, Ottawa.

SHAO, J., et RAO, J.N.K. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhya*, Series B, 55, 393-414.

WOODRUFF, R.S., et CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

numérique pour calculer les variables z_k . de Deville dans Berger et Skinner (2003)

1. INTRODUCTION

Nous remercions les trois critiques, Philip Kott, Babubhai Shah et Chris Skinner, de leurs commentaires constructifs. Nous visérons, dans notre réplique, à aborder certaines questions qu'ils ont soulevées. L'objectif principal de notre article est d'étudier l'estimation de la variance pour des estimateurs par calage de totaux de population et de paramètres non linéaires, θ , définis comme étant des solutions d'équations d'estimation « dans le cas d'un recensement ». Nous proposons une nouvelle approche de linéarisation de Taylor qui fournit un estimateur de la variance unique, en évitant de devoir évaluer initialement la statistique linéarisée aux valeurs de population. Nous montrons aussi que l'estimateur de la variance satisfait à certains critères désirables, comme l'absence de biais approximative par rapport au modèle et la validité dans des conditions d'échantillonnage répété conditionnel, du moins dans un certain nombre de cas importants. En outre, nous montrons que, dans le cas de l'échantillonnage à deux phases, l'estimateur de la variance utilise plus complètement les données de l'échantillon de première phase que les estimateurs de la variance par linéarisation classiques.

La discussion de Kott se concentre sur trois applications décrites dans notre article, à savoir i) l'estimateur de la variance par le jackknife linéarisé, v_{JL} , de l'estimateur par le quotient $\bar{Y}_R = (\bar{y}/\bar{x})$ X sous échantillonnage aléatoire simple décrit à la section 1, ii) la classe générale de poids calés par régression considérée à la section 3.2 et iii) la classe générale d'estimateurs par calage étudiée à la section 3.4. En ce qui concerne i), nous notons dans notre article le résultat selon lequel v_{JL} est à la fois asymptotiquement sans biais par rapport au plan de sondage et approximativement sans biais par rapport au modèle sous le modèle du quotient $E^m(Y_R) = \beta x_R$ et $V^m(Y_R) = \sigma^2 x_R$. Kott a raison de dire que le biais dû au modèle pourrait ne pas être négligeable si la traction d'échantillonnage, n/N , n'est pas faible. Si n/N est « petite au point d'être ignorable », alors l'absence de biais par rapport au modèle est, en fait, valide sous une fonction de variance générale $V^m(Y_R) = \sigma^2 x_R$, tel que souligné par Kott et, antérieurement, par Särndal et coll. (1989). Sous le modèle du quotient, Kott propose un estimateur de la variance plus approprié, v^m , qui est sans biais par rapport au modèle même si n/N n'est pas petite et qui est également valide sous échantillonnage répété. Les termes principaux de v^m et de v_{JL} sont identiques, et notre nouvelle approche ne reflète que le terme principal. Il

Réponse des auteurs

Il convient de souligner que l'absence de biais par rapport au modèle de v^m dépend de la validité de l'hypothèse $\sigma^2_R = \sigma^2 x_R$. En ce qui concerne ii), nous avons montré à la section 3.2 que si l'on utilise la classe générale de poids calés par régression (3.7), notre approche donne un estimateur de la variance assez complexe, faisant intervenir les moments de troisième et de quatrième ordres des poids de sondage $d_k(s)$ avec $d_k^2(s) = 0$ si le k^e élément de la population n'est pas compris dans l'échantillon s . Kott propose un choix intéressant de poids obtenus en remplaçant $c_k x_k$ dans l'expression (3.1) avec $\mathbf{q}^{(1)(k)} (\pi_k^{1/2} - \pi_k \pi_1)$ et permet aussi d'éviter les complexités liées à l'estimateur de la variance fondé sur les poids (3.7). Cette proposition est intéressante et constructive, mais $\mathbf{q}^{(1)(k)}$ nécessite la connaissance du vecteur \mathbf{x} de tous les éléments de population, contrairement à (3.7) qui dépend uniquement du total de population X ; en pratique, seul X pourrait être disponible. De surcroît, $\mathbf{q}^{(1)(k)}$ dépend de l'ensemble des $N(N-1)/2$ probabilités d'inclusion conjointe π_{kl} et, donc, le calcul de $\mathbf{q}^{(1)(k)}$ peut devenir fastidieux quand le plan d'échantillonnage est fondé sur l'échantillonnage avec probabilités inégales sans remise.

En ce qui concerne iii), Kott propose une généralisation des poids calés $w_k(s) = d_k(s) F(x_k^T \lambda)$ de la section 3.4 par ayant la même dimension que x_k . La variable z correspondante dans l'estimateur de la variance (2) est similaire à notre variable (3.21) avec $x_k x_k^T$ et $x_k y_k$ dans \mathbf{B}_T^x remplacées par $\mathbf{q}_T^x x_k^T$ et $\mathbf{q}_T^x y_k$, respectivement et $F(x_k^T \lambda)$ remplacé par $F(\mathbf{q}_T^x \lambda)$. Il s'agit d'une extension utile. Kott fait remarquer que \mathbf{B}_λ demeure un estimateur sans biais par rapport au modèle de \mathbf{B}_λ si $f(\mathbf{q}_T^x \lambda)$ dans \mathbf{B}_λ est remplacé par toute constante et que la variable z_k résultante demeure asymptotiquement inchangée sous échantillonnage répété. Cependant, Kott note aussi que le terme $f(\mathbf{q}_T^x \lambda)$ peut avoir de l'importance, même asymptotiquement, si l'on recourt au calage pour corriger la non-réponse totale en traitant la réponse d'échantillon comme un échantillonnage de deuxième phase. En utilisant le résultat pour l'échantillonnage à deux phases donné à l'annexe, Kott obtient alors un estimateur de la variance correspondant, $v(Y^{GC})$. Cette extension aux conditions de non-réponse est également utile. Il est, en effet, étonnant que le deuxième terme de $v(Y^{GC})$ donne la correction fondée sur le modèle qu'il recommande pour l'estimateur par le quotient \bar{Y}_R sous échantillonnage aléatoire simple en l'absence de non-réponse.

BIBLIOGRAPHIE ADDITIONNELLES

- DAVISON, A.C., et HINKLEY, D.V. (1997). *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.
- EFFRON, B., et HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (avec discussion). *Biometrika*, 65, 457-487.
- FULLER, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.
- WOODRUFF, R.S., et CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

substitut linéaire et à z_k , celui de variable substitut. Dans la littérature statistique plus récente, Davison et Hinkley (1997, page 46) donne à z_k le nom de valeurs d'influence empiriques. L'expression *variable linéarisée*, telle qu'utilisée par Deville (1999), me paraît simple et naturelle. Elle est conforme à l'utilisation de l'expression *statistique linéarisée* pour dénoter la statistique linéaire d'approximation et le terme *linéarisation* pour la méthode (qui est un terme général plus approprié que l'expression méthode par développement en série de Taylor pour la classe générale de méthodes considérée ici).

Un inconvénient des méthodes de linéarisation considérées ici comparativement aux méthodes de rééchantillonnage est qu'il est nécessaire d'utiliser une différenciation analytique. Il semble, d'après les exemples présentés dans l'article, que la différenciation analytique qui intervient dans la méthode proposée est au moins aussi directe que celle incluse dans les méthodes types de développement en série de Taylor des fonctions lisses de statistiques linéaires. Néanmoins, dans certaines applications, il pourrait être avantageux de remplacer le travail humain et l'erreur humaine éventuelle pouvant provenir de la différenciation analytique par l'utilisation d'une « différenciation numérique ». La méthode proposée pourrait être décrite comme étant une méthode *infinitésimale du jackknife*, puisqu'elle perturbe le poids appliqué à chaque observation d'échantillon d'une quantité infinitésimale pour déterminer la statistique linéaire d'approximation. La dérivée par rapport à un poids dans la méthode proposée pourrait être approchée numériquement par une méthode en différence finie en vertu de laquelle la statistique est recalculée en utilisant le poids perturbé par une quantité finie pour chaque observation à tour de rôle. Cette approche pourrait être décrite comme étant une méthode *jackknife* de linéarisation. Une approche classique consisterait à modifier chaque poids à tour de rôle, de sorte qu'il soit nul, peut-être en normalisant pour les poids inégaux comme dans (1.15). Il ne paraît pas essentiel de remplacer le poids original par zéro et, en principe, chaque poids pourrait être perturbé d'une certaine autre façon, par exemple, en le réduisant d'une quantité fixe δ , plus petite que la valeur minimale de $d_k(s)$. Il semble vraisemblable que, dans de nombreuses applications, l'estimateur de la variance résultant de ce genre de linéarisation jackknife aura des propriétés statistiques fort semblables à celles construites par l'approche proposée. Le choix entre les estimateurs pratiques et d'importance des calculs.

Mes derniers commentaires ont trait à la terminologie. Plusieurs raisons pratiques pourraient justifier de donner un nom à la variable z_k . En particulier, cela aiderait le praticien qui, pour certaines statistiques complexes, doit employer deux étapes de calcul distinctes, à savoir a) la construction de la variable z_k , par exemple en utilisant des routines de moindres carrés si l'on recourt à la pondération de calage et b) l'utilisation d'un logiciel type d'estimation de la variance pour les statistiques linéaires. Divers noms sont utilisés pour z_k dans la littérature. Woodruff (1971) est habituellement reconnu comme étant le premier, dans la littérature sur l'échantillonnage, à avoir attiré l'attention sur le rôle de z_k , si bien qu'Andersson et Nordberg (1994) appellent z_k la *transformation de Woodruff*. Woodruff et Causey (1976) donnent à la statistique linéaire d'approximation le nom de

intéressant de noter que la validité de la procédure proposée ne semble pas dépendre de l'exigence que $E(d(s)) = \mathbf{1}$, à condition que $\mathbf{1}$ soit remplacé par $E(d(s))$ dans le développement à la section 2. En particulier, si s représente les réponses totales et que la non-réponse peut être représentée par échantillonnage de Poisson avec probabilité de réponse inconnue, alors l'approche proposée d'estimation de la variance pourrait encore être convergente (quand elle est fondée sur de nombreux estimateurs types de la variance pour des statistiques linéaires), même si $d(s)$ est basé uniquement sur des probabilités d'inclusion d'échantillon. Julia d'Arrigo et moi-même avons récemment étudié les propriétés des estimateurs de la variance par linéarisation sous non-réponse lors d'études en simulation réalisées dans le cadre du projet de recherche DACSEIS (www.dacseis.de) en utilisant des données provenant de l'Enquête sur la population active du Royaume-Uni et de l'Enquête sur les revenus et les dépenses d'Allemagne. Nous avons considéré divers estimateurs par calage sous modèle de Poisson pour la non-réponse totale qui étaient ignorables étant donné les variables de calage, en utilisant des estimateurs types de la variance pour les statistiques linéaires sous échantillonnage stratifié à plusieurs degrés. Nous avons effectivement constaté que la non-réponse pourrait induire des biais importants dans les estimateurs de la variance par linéarisation si ceux-ci ne tiennent pas compte des poids g pour l'estimation GREG (section 3.1) ou du terme $F(x_k^g, \lambda)$ dans (3.21). Ce genre de biais est absent dans l'approche proposée.

Nous avons également étudié les estimateurs par calage de rééchantillonnage à la section 3.4. Les résultats théoriques de Deville et Samdal (1992) selon lesquels la variance asymptotique de \hat{Y}_n ne dépend pas de la forme de la fonction $F(\cdot)$ est fondée sur l'hypothèse que $\sum d_k(s) x_k$ est convergente pour X. Cette hypothèse, qui n'est pas vérifiée sous diverses sources d'erreur non dues à l'échantillonnage, n'est pas requise pour l'approche proposée. Donc, la statistique linéaire d'approximation est définie par (3.21) et l'estimateur de la variance résultant peut dépendre de la forme de $F(\cdot)$, même asymptotiquement. Les estimateurs types de la variance par linéarisation dans lesquels $d_k(s) f(x_k^g, \lambda)$ dans \hat{B}_g est remplacé par $d_k(s)$ ou par $w_k(s)$ peuvent être non convergents si ces poids diffèrent de $d_k(s) f(x_k^g, \lambda)$. Malgré ce fait théorique, nous observons peu de différences dans notre étude en simulation (pour chacune des fonctions, $1 + u, \exp(u)$ et $(1 - u)^{-1}$, utilisées pour $F(u)$) entre les propriétés statistiques des estimateurs de la variance fondés sur ces trois choix différents de poids, $d_k(s) f(x_k^g, \lambda)$, $d_k(s)$ ou $w_k(s)$, dans le vecteur \hat{B}_g dans (3.21). D'autres études pourraient produire des résultats différents.

Commentaire

CHRIS SKINNER¹

de population, \bar{X} , et qu'on suppose que cette valeur est égale à la valeur limite d'une statistique d'échantillon \bar{x} . Pour des statistiques qui sont des fonctions lisses de statistiques linéaires, il semble que l'estimateur de la variance généré par la méthode proposée puisse aussi être construit par les méthodes classiques de développement en série de Taylor, à condition qu'aucune simplification initiale de l'estimateur de la variance ne soit faite en se fondant sur ce genre d'hypothèse au sujet de l'information auxiliaire. Une telle construction pourrait, cependant, être moins précise que l'approche proposée.

Les hypothèses associées aux méthodes de linéarisation qui diffèrent de celles de l'approche proposée, comme celle selon laquelle une valeur d'échantillon \bar{x} est la valeur théorique limite d'une valeur d'échantillon \bar{x} , sont fondées sur des lois inconditionnelles et, par conséquent, on pourrait s'attendre à ce que l'intégration de ce genre d'hypothèse dans un estimateur de la variance puisse endommager les propriétés conditionnelles de la méthode, particulièrement en ce qui a trait à des statistiques telles que x . La procédure proposée évite de dépendre de ce genre d'hypothèse et, grâce au calcul des dérivés pour des valeurs réalisées d'échantillon, devrait, en principe, pouvoir suivre les propriétés conditionnelles de plus près. (Il semble exister un parallèle avec les arguments d'Efron et de Hinkley (1978) en faveur de l'information observée par opposition à l'information attendue, quoique le contexte soit assez différent.)

Éviter de dépendre de ce genre d'hypothèse pourrait non seulement améliorer les propriétés conditionnelles de l'approche proposée, mais aussi protéger l'estimateur de la variance contre les effets éventuellement biaisants des erreurs non dues à l'échantillonnage. L'information de population auxiliaire pourrait différer des valeurs limites de la statistique d'échantillon correspondante, à cause de la non-réponse ou de la non-couverture, ou à cause de divergences dans la façon dont les variables auxiliaires sont mesurées. Dans de telles circonstances, les méthodes de linéarisation qui diffèrent de l'approche proposée pourraient donner lieu à une estimation de la variance non convergente. Pour cette raison, Fuller (2002, page 10) recommande d'utiliser les poids g dans (3.6), tel qu'il est proposé, particulièrement en présence de non-réponse (page 15). En ce qui concerne le dernier cas, il semble

Les approches de linéarisation et de rééchantillonnage fournissent deux classes générales de méthodes d'estimation de la variance dans les enquêtes. Toutes deux ont des avantages relatifs et il semble important de leur garder à l'enquête. Le présent article approfondit notre compréhension générale pour générer ce genre d'estimateur de la variance de façon unique et fournit des exemples utiles de son application dans certains domaines importants.

Une méthode de linéarisation consiste à approximer la variance d'une statistique d'intérêt par la variance d'une statistique linéaire pour laquelle on suppose qu'il existe un estimateur de la variance approprié. La question principale ici est la méthode utilisée pour déterminer la statistique linéaire. La méthode type suppose que la statistique d'intérêt peut être exprimée sous forme d'une fonction dérivable d'un vecteur de statistiques linéaires (de dimension fixe) et utilise le développement en série de Taylor pour déterminer l'approximation. L'approche proposée dans le présent article s'applique à une classe plus générale de statistiques pondérées d'après l'échantillon, illustrées par les exemples complexes des sections 3.2 et 4. L'estimateur de la variance est construit par différenciation de la statistique par rapport aux poids d'échantillonnage. L'approche de l'approximation linéaire est étroitement liée aux méthodes fondées sur la fonction d'influence (par exemple, les équations 1.6 et 1.13) et l'article fournit une revue utile de ce genre de méthodes à la section 1. Les auteurs notent qu'il n'est pas facile de vérifier la validité de ces méthodes pour les statistiques qui ne sont pas des fonctions lisses de (un nombre fixe de) statistiques linéaires et il serait intéressant de savoir à quel point l'approche proposée fournit effectivement des estimateurs de la variance valides pour des statistiques, comme les quantiles, qui ne sont pas de cette forme.

Une caractéristique importante de l'approche proposée, qui assure la construction unique de l'estimateur de la variance, est que les dérivés sont calculés à des valeurs fondées sur l'échantillon réalisés, sans aucune évaluation initiale de la statistique d'approximation linéaire aux valeurs théoriques de population. Ce genre d'évaluation initiale pourrait effacer le caractère unique, par exemple, si de l'information auxiliaire est disponible sur une moyenne

Commentaire

BABUBHAI V. SHAH

Il s'agit d'un article excellent qui estompe le mystère qui sous-tend la linéarisation de Taylor. La plupart des applications d'analyse de données font usage des poids d'Horvitz-Thompson qui sont les inverses des probabilités de sélection. La prescription la plus simple pour calculer la linéarisation pour un estimateur $\hat{\theta}$ est la suivante :

- 1. Pour chaque observation, créer une nouvelle variable $z_i = \partial/\partial w_i$, où w_i est l'inverse de la probabilité de sélection pour la i^{e} observation sélectionnée dans l'échantillon. Quand l'estimateur $\hat{\theta}$ est défini implicitement au moyen d'équations d'estimation, la dérivée peut être calculée par différenciation des équations implicites.

2. Définir le total pondéré $T' = \sum w_i z_i$.

- 3. Calculer la variance V' du total T' basée sur le plan de sondage.

- 4. La variance V' est la variance approximative de l'estimateur $\hat{\theta}$.

Si le paramètre θ est un vecteur, alors la variable z_i et le total T' sont aussi des vecteurs et V' est une estimation approximative de la matrice des variances-covariances de l'estimateur $\hat{\theta}$.

Les étapes (1) et (2) susmentionnées produisent la linéarisation correcte dans les cas suivants :

- a. moyennes, proportions et estimations de ratio;
- b. modèles de régression linéaire généralisés;
- c. valeur marginale prédite pour un modèle linéaire généralisé;
- d. estimation de la moyenne d'après des données imputées par régression;
- e. modèles de régression linéaire généralisée avec poids calés;

- f. test de la somme des rangs pour deux échantillons de Wilcoxon;
- g. estimations des coefficients et du taux de risque dans le modèle à risques proportionnels de Cox;
- h. estimations de la survie marginale prédite dans le modèle à risques proportionnels de Cox;
- i. enquête par échantillonnage à deux phases.

À l'étape (1), la dérivation est définie de façon unique, ne contient pas la vraie valeur du paramètre θ , et ne nécessite pas le remplacement par l'estimateur $\hat{\theta}$.

L'indépendance de l'étape (3) pour le calcul de la variance à partir de la linéarisation aux étapes (1) et (2) est bien démontrée par la discussion sur l'échantillonnage à deux phases de la section 4. Dans la plupart des cas, on suppose qu'on a fait à un plan de sondage avec remise pour estimer la variance du total à l'étape (3). Naturellement, on pourrait obtenir une meilleure estimation de la variance du total en utilisant toute l'information disponible sur le plan de sondage. Pour un plan d'échantillonnage à deux phases, on peut exécuter l'étape (1) en utilisant les poids d'Horvitz-Thompson pour la première phase et en traitant les multiplicateurs m_i comme des données. Le multiplicateur m_i est nul si l'observation i n'est pas sélectionnée à la deuxième phase et est égal à l'inverse de la probabilité conditionnelle $\pi_{2k/1}^i$. L'étape résultante (2) produit le même total que celui présenté au paragraphe précédent, à la section 4, décrit le moyen approprié d'estimer la variance de ce total pour un plan d'échantillonnage à deux phases sans remise à chaque phase, et ce calcul est indépendant de la linéarisation.

Les étapes (1) et (2) génèrent la linéarisation appropriée dans tous les cas connus, sauf celui, comme le quantifie, où l'estimateur n'est pas une fonction continue des poids w_i .

Permettez-moi de conclure ces remarques en remerciant MM. Dennati et Rao de leur article stimulant, ainsi que *Techniques d'enquête* de l'avoir publié et de m'avoir permis de faire certains commentaires.

BIBLIOGRAPHIE ADDITIONNELLE

- ESTEVAO, V.M., et SÄRNDAAL, C-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.
- FULLER, W.A., LOUGHIN, M.M. et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.
- GODAMBE, V.P., et THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *International Statistical Review*, 54, 2, 127-138.
- KOTT, P.S., et BREWER K.R.W. (2001). Estimating the model variance of a randomization-consistent regression estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 81 1-822.
- TILLÉ, Y. (1999). Estimation dans des enquêtes par sondage avec des probabilités d'inclusion conditionnelles: enquêtes à plan d'échantillonnage complexe. *Techniques d'enquête*, 25, 61-71.

où \mathbf{q}^k à de nouveau la même dimension que \mathbf{x}^k . Par souci de commodité, on suppose que F est positive et dérivable de deux fois au voisinage de $\mathbf{q}_T^k \lambda$. Sans perte de généralité, on peut supposer que λ (la limite de λ_j) est 0, et que $f(0) > 1$. Si $Y^{GC} = \sum^U w_k^{(s)} y_k$ est un estimateur converti sous randomisation, et je suppose qu'il l'est, $f(0)$ est égale à 1. Un développement parallèle à celui présenté dans le texte mène en dernière analyse à

$$z_k = F(\mathbf{q}_T^k \lambda) (y_k - x_k^T \mathbf{b}^k) = F(\mathbf{q}_T^k \lambda) e_{k\lambda}$$

Ceci diffère de l'estimateur de la variance présenté dans Folsom et Singh (2000), principalement parce que ces auteurs supposent que l'échantillon original est sélectionné selon un plan d'échantillonnage stratifié à plusieurs degrés avec remise au premier degrés. Cela, entre autres choses, annule la deuxième sommation dans le deuxième membre. Non seulement $V(Y^{GC})$ estime l'erreur quadratique moyenne sous quasi-randomisation de Y^{GC} – « quasi » parce qu'on émet l'hypothèse d'un modèle de réponse – mais il estime aussi la variance fondée sur un modèle de Y^{GC} . En fait, le biais relatif de $V(Y^{GC})$ sous le modèle de prédiction, $E_{m(y)}(x^k | \mathbf{q}^k) = x_k^k \mathbf{b}$, est $O(1/n)$ quand les y_k ne sont pas corrélés et $V_{m(y)}(x^k | \mathbf{q}^k) = x_k^k \gamma$, où γ (comme \mathbf{b}) ne doit pas être spécifié. Etonnamment, le deuxième terme de $V(Y^{GC})$ fournit la correction fondée sur le modèle que j'ai recommandée pour l'estimateur par le quotient sous échantillonnage aléatoire simple en l'absence de non-réponse.

L'estimateur de la variance « par remplacement fonctionnel-t-il réellement pour la forme d'Horvitz-Thompson complète?

Comme je l'ai mentionné entre parenthèses au début, j'ai omis la phrase cle « sous des conditions faibles que je suppose être vérifiées » à plusieurs reprises dans ces commentaires. Maintenant, je veux examiner ce que pourrait être une de ces conditions. Il est typique, en estimation (ou de modèle) par les valeurs d'échantillon correspondantes, puis que les écarts sont asymptotiquement ignorables. C'est ce que font, par exemple, Demnati et Rao dans l'équation (2.4) quand ils remplacent z_k par z_k . La question que j'aimerais soulever, et dont je ne connais pas la réponse, est la suivante. Supposons qu'on estime un total $O(N)$ par un estimateur par calage. Le total est $O(N)$, et $O(n) = O(N)$. La variance fondée sur le modèle et l'erreur quadratique moyenne sous randomisation de l'estimateur sont aussi $O(n)$. Est-il légitime de remplacer z_k par z_k où $z_k - z_k = O^p(1/\sqrt{n})$, quand il existe $n(n-1)/2$ termes dans l'estimateur de la variance/erreur quadratique moyenne d'Horvitz-Thompson ou de Yates-Grundy? Dans la plupart des applications pratiques la question ne se pose pas, parce que l'estimateur de la variance peut être réexprimé avec

où $\mathbf{b}^k = [\sum^d d_k(s) f(\mathbf{q}_T^k \lambda) \mathbf{q}^k x_T^{T-1} \sum^d d_k(s) f(\mathbf{q}_T^k \lambda) \mathbf{q}^k y_k^k]$. La présence de $f(\cdot)$ dans l'expression de \mathbf{b}^k est peut-être une surprise, mais, comme il s'avère, elle n'est pas significative dans ce contexte. Pour l'inférence sous le modèle de prédiction, $E_{m(y)}(x^k | \mathbf{q}^k) = x_k^k \mathbf{b}$, la dérivée peut être remplacée par toute constante sans conséquence asymptotique; \mathbf{b}^k demeure un estimateur sans biais fondé sur le modèle pour \mathbf{b} . Pour l'inférence fondée sur la randomisation, puisque $\mathbf{q}_T^k \lambda = O^p(1/\sqrt{n})$ et $f(0) > 0$, z_k ne serait pas affecté asymptotiquement si on remplaçait $f(\mathbf{q}_T^k \lambda)$ par 1 ou par $F(\mathbf{q}_T^k \lambda)$.

Par contre, les choses changent si nous approfondissons un peu les choses. Fuller, Loughin et Baker (1994) ont utilisé le calage pour corriger pour la non-réponse totale en traitant la réponse d'échantillon comme une deuxième phase d'échantillonnage. Il suppose que chaque élément k de la population a une probabilité de Poisson de réponse d'échantillon, π_k , qui ne dépend pas du fait qu'il soit effectivement sélectionné dans l'échantillon ou non. Ils supposent en outre que $\pi_k = 1/(1 + x_T^k \lambda)$, où λ est inconnu et estimé implicitement par calage. Ici, nous généralisons cela et supposons que $\pi_k = 1/F(\mathbf{q}_T^k \lambda)$, où F est connue, positive et dérivable deux fois. En pratique, \mathbf{q}^k sera vraisemblablement identique à \mathbf{x}^k , mais il peut être raisonnable de remplacer une ou plusieurs composantes de \mathbf{x}^k par des variables que l'on suppose être plus fortement corrélées avec la réponse/non-réponse.

En redéfinissant s comme étant l'échantillon répondant et $d_k(s)$ comme étant $(1/\pi_k)$ quand $k \in s$, 0 autrement, tout se passe comme avant. La différence est que $f(\mathbf{q}_T^k \lambda)$ dans \mathbf{b}^k ne doit plus être asymptotiquement identique sur les k . Donc, le terme peut avoir de l'importance même dans le cas d'un grand échantillon. Maintenant $V(Y^{GC}) \approx V(\sum^U d_k(s) z_k)$, où $\sum^U d_k(s) z_k = \sum^U d_k(s) F(\mathbf{q}_T^k \lambda) e_{k\lambda}$ est l'estimation par double expansion. En substituant $1/F(\mathbf{q}_T^k \lambda)$ à π_k , l'estimateur de la variance de Y^{GC} devient (d'après l'équation (A.1) avec $\pi_{2kj/1} = \pi_{2kj} \pi_{2k} \pi_{2j}$)

$O(n)$ termes. Qu'en est-il si ce n'est pas le cas?

Commentaire

Dermati et Rao : Estimateurs de la variance par linéarisation pour des données d'enquête

L'article traite d'un nombre impressionnant de contextes, dont bon nombre n'ont été étudiés que récemment dans la littérature, souvent par le professeur Rao lui-même. Je n'ai que peu de chose à dire ici au sujet des fonctions d'estimation avec poids calés ou de l'échantillonnage à deux phases, si ce n'est (principalement) que je suis d'accord avec les solutions préconisées dans le texte. Je me concentrerai plutôt sur trois applications, à savoir l'estimateur par le quotient sous échantillonnage aléatoire simple discuté dans l'introduction, la classe générale de poids calés par régression décrite à la section 3.2 et la classe générale d'estimateurs par calage décrits à la section 3.4. Je terminerai par une question sur l'estimateur de la variance par linéarisation sous la forme d'Horvitz-Thompson complète qui me tracasse depuis un certain temps.

d'utiliser des poids calés de la forme

Une classe générale d'estimateurs par calage

Puisque $q^{(2)}$ est une fonction de l'échantillon, les auteurs nous entraînent dans les complications de la section 3.2. Cela n'est nécessaire que pour l'inférence fondée sur la réadomposition. J'aurais choisi une autre voie. Observons que $b(s)^{(2)}p^{(2)}b(s)^{(1)} = O(1/n)$. Le remplacement de a par un \hat{a} et a asymptotiquement ignorable sur $w^{(s)}$ (autrement dit, la différence relative est $O(1/n)$).

$$\sum_{j \in S} (u^{k_j} u^j - x^j u^{k_j} u^j) / (u^{k_j} u^j) = {}^{(2)}k b$$

Un autre choix, étudié indirectement par Demnati et Rao, représente les propriétés fondées sur la randomisation. et aboutissant aussi à une variante de l'estimateur optimal sous randomisation, est

Un choix pour \mathbf{b} est
$$\sum_{\ell \in f} \frac{\mathbf{x}^{\ell} \mathbf{x}^{\ell} / \mathbf{x}^{\ell} \mathbf{x}^{\ell}}{\mathbf{x}^{\ell} \mathbf{x}^{\ell} / \mathbf{x}^{\ell} \mathbf{x}^{\ell}} = \mathbf{b}$$
 dont l'utilisation aboutit à une variante de l'estimateur par régressions optimal sous randomisation proposée par Tibshirani (1996). Observons que $(\sum_{\ell \in f} \mathbf{x}^{\ell} \mathbf{x}^{\ell})^{-1} \text{Cov}(\mathbf{X}, Y)$, où Var et Cov

$$(\pi_{\mathcal{U}}^f)^{-1} \pi_{\mathcal{V}}^f / (\pi_{\mathcal{U}}^f)^{-1} \pi_{\mathcal{W}}^f = \sum_{j \in U} \pi_{\mathcal{V}}^f \pi_{\mathcal{W}}^f = \pi_{\mathcal{V}}^f \pi_{\mathcal{W}}^f$$

Je généraliserai personnellement les résultats de la section 3.1 d'une autre façon que celle décrite par les auteurs à la remplacection 3.2. À l'exemple d'Estavao et Samdal (2002), je remplacerais $c_k x_k$ dans l'équation (3.1) par un vecteur \mathbf{b}_k ayant la même dimension que x_k . Le reste de la section se déduit facilement.

Une classe générale de poids calés par régression

Quand on utilise l'échantillonnage aléatoire simple en $O(1/\sqrt{n})$, tout comme ν_L et ν_T .

de la variance de \hat{Y}^m est $V^m = \frac{1}{n} \left(\frac{X}{X-1} \right)^2 \left[1 - \frac{n}{N} \right] \left[\frac{X}{X-1} \right]$ (Kott et Brewer, 2001). Comme un estimateur de l'erreur quadratique moyenne sous randomisation de \hat{Y}^R , V^m a un biais relatif de

et $Cov^2(Y^{(l)}(y), X^{(l)}(x))$ nous obtenons

$$Cov[Y^{(l)}(y), X^{(l)}(x)] = Cov[X^{(l)}(y), X^{(l)}(x)].$$

Un estimateur sans biais de type H-T de $2Cov[X^{(l)}(y), X^{(l)}(x)]$ est donné par

$$2cov[X^{(l)}(y), X^{(l)}(x)]$$
$$= 2 \sum_{k=1}^{k_{les}} \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{kl}^*} \frac{\pi_{1k} \pi_{1l}}{Y_k X_l}.$$

(A.3)

La somme de (A.1), (A.2) et (A.3) est égales (4.5).

BIBLIOGRAPHIE

ANDERSON, C., et NORDBERG, L. (1994). A method for variance estimation of non-linear functions of totals in surveys - theory and software implementation. *Journal of Official Statistics*, 10, 395-405

BERGER, Y.G. (2002). A generalized jackknife variance estimator for nonlinear statistics in probability sampling. Rapport technique, Department of Social Statistics, University of Southampton.

BINDER, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette» *Techniques d'enquête*, 22, 17-22.

CAMPBELL, C. (1980). A different view of finite population estimation. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 319-324.

DEVILLE, J.C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes: linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

DEVILLE, J.C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. et STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons, Inc.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons, Inc.

KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

MONTANARI, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 71-79.

RAO, J.N.K., et SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

RAO, J.N.K., YUNG, W. et HINDRIGLOU, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sanhya*.

ROYAL, R.M., et CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J.H. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SERFLING, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons, Inc.

SINGH, A.C., et FOLSOM, R.E. (2000). Bias correcting estimating function approach for variance estimation adjusted for poststratification. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 610-615.

SITTER, R.R., et WU, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association*, 97, 535-544.

SUKHATME, P.V., et SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. 2^{ème} éd. London: Asia Publishing House.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

YUNG, W., et RAO, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.

et

$$z_k = \frac{\bar{X}_{(1)}}{\bar{X}_{(1)}}(y_k - \bar{R}x_k) = \frac{\bar{X}}{\bar{X}_{(1)}}e_k \quad (4.6)$$

$$z_k^{(1)} = \bar{R}x_k^{(1)} \quad (4.7)$$

Sous échantillonnage aléatoire simple aux deux phases, (4.6) et (4.7) se réduisent à $z_k^{(1)} = (\bar{x}/x)x_k^{(1)}$ et $z_k^{(1)} = (\bar{y}/y)y_k^{(1)}$ où $e_k = y_k - (\bar{y}/x)x_k^{(1)}$ et \bar{x} sont les moyennes d'échantillon de deuxième phase de y et x respectivement, et $\bar{x}_{(1)}$ est la moyenne d'échantillon de première phase de x . Maintenant, en substituant z_k et $z_k^{(1)}$ à y et x dans (4.5) et en notant que $\pi_{1kl} = \pi(n-1)/[N(N-1)]$, $\pi_{2kl/1} = \pi(n-1)/[n(n-1)]$, $\pi_{1k} = \pi^{1k}$ et $\pi_{2kl/1} = \pi^{2kl/1}$, nous obtenons

$$v_L(\bar{Y}^{R2}) = N^2 \left(\frac{1}{1} - \frac{N}{1} \right) \bar{R}^2 s_{1x}^2 + N^2 \left(\frac{1}{1} - \frac{N}{1} \right) \bar{R} \frac{\bar{x}}{s_{ex}} \quad (4.8)$$

où

$$\bar{R} = \bar{y}/\bar{x}, s_{1x}^2 = (n-1)^{-1} \sum_{k \in s_1} (x_k - \bar{x}_{(1)})^2,$$

$$s_{2e}^2 = (m-1)^{-1} \sum_{k \in s} (e_k - \bar{e})^2,$$

$$s_{2ex}^2 = (m-1)^{-1} \sum_{k \in s} (e_k - \bar{e})(x_k - \bar{x})$$

et \bar{e} est la moyenne d'échantillon de deuxième phase de e . La formule (4.8) concorde avec la formule établie par Rao et Sitter (1995). Elle diffère de la formule habituelle (Sukhatme et Sukhatme 1970, page 176) qui omet d'utiliser l'entité des données $x \{x_k, k \in s_1\}$. Rao et Sitter (1995) ont démontré par simulation que $v_L(\bar{Y}^{R2})$ est plus efficace que l'estimateur de la variance habituel. En outre, $v_L(\bar{Y}^{R2})$ donne de meilleurs résultats en ce qui concerne le suivi de l'erreur quadratique moyenne conditionnelle de \bar{Y}^{R2} ; voir Rao et Sitter (1995, section 3) pour des précisions sur l'étude en simulation.

CONCLUSION

Nous avons présenté une approche unifiée de calcul d'estimateurs de la variance par linéarisation de Taylor et nous l'avons appliquée à divers problèmes. Elle mène directement à un estimateur de la variance présentant certaines propriétés désirables, du moins dans un certain nombre de cas particuliers importants, notamment l'absence d'approximative de biais pour la variance fondée sur un modèle de l'estimateur sous un modèle hypothétique et la validité sous l'échantillonnage répété conditionnel. Il serait utile de déterminer si ce genre de propriétés désirables

tiennent aussi pour des cas plus complexes, comme la classe générale d'estimateurs par calage (section 3.2), les estimateurs fondés sur les équations d'estimation (section 3.3) et l'échantillonnage à deux phases (section 4). À l'heure actuelle, nous étudions diverses extensions de notre méthode, y compris l'estimation de la variance sous imputation pour la non-réponse partielle et l'estimation de la variance à partir de données d'enquête longitudinales.

REMERCIEMENTS

Nous remercions le rédacteur adjoint et un examinateur pour leurs suggestions et commentaires constructifs. Nous remercions aussi plusieurs collègues à Statistique Canada de leurs suggestions utiles et de leurs encouragements, particulièrement Linda Standish, David Binder, Geoff Hole, Richard Burgess et Larry Swain. Les travaux de Demati ont été financés par la Division des données régionales et administratives de Statistique Canada. Ceux de J.N.K. Rao ont été financés par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

ANNEXE

Estimateur de la variance sans biais de $\bar{Y}_{(1)}$ et $\bar{X}_{(1)}$ (x)

La variance de $\bar{Y}_{(1)}$ et $\bar{X}_{(1)}$ est la somme de la variance de $\bar{Y}_{(1)}$ et de la variance de $\bar{X}_{(1)}$ et de deux fois la covariance de $\bar{Y}_{(1)}$ et $\bar{X}_{(1)}$. Un estimateur sans biais de type H-T de $V[\bar{Y}_{(1)}]$ est donné par Stådal, Swensson et Wretman (1991, chapitre 9, page 348) :

$$v[\bar{Y}_{(1)}] = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{1kl}^*}{\pi_{1k} - \pi_{1l}} \frac{\pi_{1k}}{y_k} \frac{\pi_{1l}}{y_l} + \sum_{k \in s} \sum_{l \in s} \frac{\pi_{2kl/1}^*}{\pi_{2k/1} - \pi_{2l/1}} \frac{\pi_{2k/1}}{y_k} \frac{\pi_{2l/1}}{y_l} \quad (A.1)$$

Un estimateur sans biais de type H-T de $V[\bar{X}_{(1)}]$ est donné par

$$v[\bar{X}_{(1)}] = \sum_{k \in s_1} \sum_{l \in s_1} \frac{\pi_{1kl}^*}{\pi_{1k} - \pi_{1l}} \frac{\pi_{1k}}{x_k} \frac{\pi_{1l}}{x_l} + \text{Cov}[\bar{Y}_{(1)}, \bar{X}_{(1)}] = E\text{Cov}[\bar{Y}_{(1)}, \bar{X}_{(1)}] = E\text{Cov}[\bar{Y}_{(1)}, \bar{X}_{(1)}] \quad (A.2)$$

En outre, où E_2 et Cov_2 représentent l'espérance conditionnelle et la covariance conditionnelle sachant s_1 . En notant que

$$E_2 \bar{Y}_{(1)} = \bar{X}_{(1)}(y), E_2 \bar{X}_{(1)} = \bar{X}_{(1)}(x)$$

Nous prenons maintenant la dérivée de (3.14) par rapport à b_k pour obtenir

$$\sum_{N=1}^N [\partial(b_l g^k(b)) / \partial b_k] [u^l(\theta(b)) + \sum_{N=1}^N (b_l g^k(b)) / \partial b_k] + \sum_{N=1}^N (b_l g^k(b)) / \partial b_k$$

$$(3.15) \quad \left[\partial u^l(\theta(b)) / \partial(\theta(b)) \right] \partial(\theta(b)) / \partial b_k.$$

En substituant (3.2) et (3.3) à $\partial(b_l g^k(b)) / \partial b_k$ dans (3.15), nous obtenons (3.13) après simplification. Ce résultat montre que notre méthode est également directement applicable aux estimateurs généraux θ sous les conditions de régularité de Binder (1983).

3.4 Une classe générale d'estimateurs par calage

Les poids calés, $w_k(s)$, associés à l'estimateur GREG \hat{Y}^w pourraient ne pas être systématiquement non négatifs.

Pour contourner cette difficulté, on utilise souvent des poids généralisés calés par la méthode itérative du quotient (raking ratio). Ces poids sont toujours non négatifs, mais la méthode peut produire certaines valeurs extrêmes (Deville et Särndal 1992).

Les poids généralisés par la méthode itérative du

$$(3.16) \quad w_k^s(s) = d_k^s(s) F(x_T^k \lambda)$$

avec $F(a) = e^a$, où le multiplicateur de Lagrange λ est déterminé en résolvant les équations de calage

$$(3.17) \quad \sum w_k^s(s) x_k = \sum d_k^s(s) F(x_T^k \lambda) x_k = X.$$

Les poids GREG correspondent à $F(a) = 1 + a$, auquel cas

$$\hat{\lambda} = (\sum d_k^s(s) x_k x_T^k)^{-1} (X - X).$$

En général, l'estimateur par calage $\hat{Y}^w = \sum w_k^s(s) y_k$ avec les poids $w_k^s(s)$ donnés par (3.16) pourrait ne pas être exprimable sous la forme d'une fonction de totaux estimés. Par conséquent, nous suivons l'approche de Binder (1983) et étendons $F(x_T^k \lambda)$ autour de λ , où λ représente la limite de probabilité de $\hat{\lambda}$. Nous obtenons

$$(3.18) \quad F(x_T^k \lambda) \approx F(x_T^k \lambda) + f(x_T^k \lambda) x_T^k (\hat{\lambda} - \lambda),$$

où $f(a) = \partial F(a) / \partial a$. En outre, en étendant les équations de calage (3.17) autour de λ , nous obtenons après simplification

$$(3.19) \quad \hat{\lambda} - \lambda \approx -\hat{\mathcal{Q}}_{\lambda}^{-1} (\hat{S}_{\lambda} - X)$$

où $\hat{\mathcal{Q}}_{\lambda} = \sum d_k^s(s) f(x_T^k \lambda) x_k x_T^k$ et $\hat{S}_{\lambda} = \sum d_k^s(s) F(x_T^k \lambda) x_k$. Notons que $\hat{\mathcal{Q}}_{\lambda}$ ainsi que \hat{S}_{λ} sont de la forme de totaux estimés. En substituant (3.19) dans (3.18), nous obtenons

$$(3.20) \quad F(x_T^k \lambda) \approx F(x_T^k \lambda) - f(x_T^k \lambda) x_T^k \hat{\mathcal{Q}}_{\lambda}^{-1} (\hat{S}_{\lambda} - X).$$

En utilisant l'approximation (3.20) dans (3.16), il s'ensuit que \hat{Y}^w est approximé par une fonction différentiable de totaux estimés. Donc, la théorie générale de la section 2 est

appliquable et il reste à évaluer $z_k^* = \partial h(b) / \partial b_k |_{b=d(s)}$, où

$$h(b) = \sum b_k g_k^*(b) \text{ avec}$$

$$g_k^*(b) = F(x_T^k \lambda) - f(x_T^k \lambda) x_T^k \hat{\mathcal{Q}}_{\lambda}^{-1} (S_{\lambda}(b) - X)$$

$$(3.21) \quad z_k^* = F(x_T^k \lambda) - x_T^k \hat{B}_{\lambda}^* = F(x_T^k \lambda) e_{k\lambda},$$

$$\hat{B}_{\lambda}^* = \sum d_k^s(s) f(x_T^k \lambda) x_k x_T^k$$

Singh et Folsom (2000) obtiennent un résultat comparable en utilisant une approche légèrement différente.

Le résultat (3.21) peut aussi être obtenu directement par la même voie que (3.2) et (3.3) en écrivant \hat{Y}^w sous la forme $f(d(s))$ et en évaluant $z_k^* = \partial f(b) / \partial b_k |_{b=d(s)}$, où $f(b) = \sum b_k g_k^*(b) y_k$ avec $g_k^*(b) = F(x_T^k \lambda)$. Nous avons

$$(3.22) \quad \partial(b_k g_k^*(b)) / \partial b_k = g_k^*(b) + b_k f(x_T^k \lambda(b)) x_T^k \partial \lambda(b) / \partial b_k$$

et, pour $l \neq k$

$$(3.23) \quad \partial(b_l g_l^*(b)) / \partial b_k = b_l f(x_T^l \lambda(b)) x_T^l \partial \lambda(b) / \partial b_k.$$

Pour évaluer $\partial \lambda(b) / \partial b_k$, nous prenons les dérivées des équations de calage (3.17) avec $d(s)$ remplacé par b :

$$\sum b_k^* F(x_T^k \lambda(b)) x_k = 0. \text{ Ceci nous donne}$$

$$0 = F(x_T^k \lambda(b)) x_k + \sum_l b_l f(x_T^l \lambda(b)) x_T^l \partial \lambda(b) / \partial b_k$$

$$(3.24) \quad \partial \lambda(b) / \partial b_k = - \left(\sum_l b_l f(x_T^l \lambda(b)) x_T^l \right)^{-1} F(x_T^k \lambda(b)) x_k.$$

Par substitution de (3.24) dans (3.22) et (3.23), nous obtenons (3.21) après simplification.

Deville et Särndal (1992) montrent que la variance asymptotique de \hat{Y}^w pour une $F(\cdot)$ générale est équivalente à la variance asymptotique de l'estimateur GREG qui comprend le coefficient de régression « de recensement » B . En utilisant ce résultat, ils obtiennent un estimateur de la variance de \hat{Y}^w pour une $F(\cdot)$ générale, en remplaçant B par $\hat{B} = (\sum w_k^s(s) x_k x_T^k)^{-1} \sum w_k^s(s) y_k$, où $w_k^s(s) = d_k^s(s) F(x_T^k \lambda)$. La variable z_k^* résultante concorde avec notre donnée par (3.21) si $f(a) = F(a)$, c'est-à-dire dans le cas des poids obtenus par ajustement itératif généralisé. Dans le cas de l'estimateur GREG, nous avons $f(x) = 1 + x$, $f(x) = 1$ et \hat{B} se réduit à $g_k(\hat{b}(s))$ se réduit au poids $g_k(\hat{b}(s))$ et que $e_{k\lambda} = y_k - x_T^k \hat{B}_{\lambda}^*$ avec $\hat{B}_{\lambda}^* = (\sum d_k^s(s) x_k x_T^k)^{-1} \sum d_k^s(s) x_k y_k$. Notons que, dans ce cas, notre variable z_k^* est différente de celle de Deville et Särndal (1992), mais concorde avec une variable z_k^* utilisée couramment (Särndal, Swensson et Wretling 1989).

comprend les moments de troisième et de quatrième ordres $E[d_k(s)d_l(s)d_q(s)]$ et $E[d_k(s)d_l(s)d_q(s)d_p(s)]$ en plus des moments de deuxième ordre $E[d_k(s)d_l(s)]$, tandis que l'estimateur de la variance des estimateurs par la régression généralisée ne nécessite que les deuxième moments. En particulier, si $d_k(s) = (1/\pi_l)a_k(s)$, nous avons besoin des probabilités d'inclusion de troisième et de quatrième ordres π_{klq} et π_{klqp} , ainsi que les probabilités d'inclusion de deuxième ordre π_{kl} .

Le calcul de z_k et z_{kl} comprend les dérivées $\partial[b_l h(\theta_{(1)}), b_{(2)}] / \partial b_k$ pour $l \neq k$ et les dérivées $\partial[b_l h(\theta_{(1)}), b_{(2)}] / \partial b_l$ pour $l = k$ et $l \neq k$. Après simplification, nous obtenons

$$z_k = \left[1 + (X - X)^T \tilde{Q}^{-1} c_k x_k \right] e_k^*$$

et

$$z_{kl} = (X - X)^T \tilde{Q}^{-1} c_{kl} x_l e_k^*$$

où

$$e_k^* = y_k - x_k^T \beta^*$$

avec $\beta^* = \tilde{Q}^{-1} (\sum_k d_k(s) c_k x_k y_k + \sum_{l \neq k} d_l(s) d_k(s) c_{kl} x_l y_k)$.

Notons que l'estimation de la variance par linéarisation de Taylor habituelle comprend l'utilisation de $v(e^*)$, tandis que $v(z_{(1)}, z_{(2)})$ ferait intervenir les résidus e_k^* , ainsi que les poids g_k et $(X - X)^T \tilde{Q}^{-1} c_k x_k$ et $(X - X)^T \tilde{Q}^{-1} c_{kl} x_l$. Si $c_{kl} = 0$ pour tout $k \neq l$, alors $z_{kl} = 0$ et $v(z_{(1)}, z_{(2)})$ se réduit à $v(z)$ avec z_k donné par (3.6). Donc, le résultat GREG de la sous-section 3.1 est un cas spécial.

3.3 Equations d'estimation

Nous examinons maintenant le paramètre vectoriel

$\theta = (\theta_1, \dots, \theta_p)^T$ défini explicitement ou implicitement comme étant de solution à "recensement" d'équations d'estimation $S(\theta) = \sum_{k=1}^N n_k^k(\theta) = 0$. On obtient une estimation par calage $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^T$ avec poids calés GREG $w_k^k(s) = d_k^k(s) g_k^k(p(s))$ par la résolution des équations d'estimation sur l'échantillon :

$$S(\hat{\theta}) = \sum w_k^k(s) n_k^k(\hat{\theta}) = 0, \quad (3.10)$$

où $n_k^k(\hat{\theta})$ et $S(\hat{\theta})$ sont des vecteurs de dimension $(p \times 1)$ (Binder 1983). Par exemple, pour la régression logistique avec le scalaire θ , nous avons $n_k^k(\hat{\theta}) = (y_k - p_k^k(\hat{\theta})) a_k^k$ où $p_k^k(\hat{\theta}) = P(y_k = 1 | a_k^k) = \exp(\theta a_k^k) / (1 + \exp(\theta a_k^k))$ et a_k^k est la variable prédictive. Notons que dans ce cas, $\hat{\theta}$ est la solution implicite de (3.10) et est obtenu itérativement par la méthode des scores de Newton-Raphson ou de Fisher.

L'estimateur d'un ratio de totaux Y et de $A = \sum a_k^k$ est obtenu en tant que solution explicite de (3.10) avec $n_k^k(\hat{\theta}) = y_k - \theta a_k^k$; $\hat{\theta} = \sum w_k^k(s) y_k / \sum w_k^k(s) a_k^k = Y/A$. Dans ce cas, $\hat{\theta}$ est une fonction des totaux estimés et, donc, notre méthode pour les fonctions de totaux est applicable. Il reste

à évaluer $\partial f(\theta) / \partial b_k$ ou $f(\theta) = \sum_{k=1}^N b_k g_k^k(y_k) / \sum_{k=1}^N b_k g_k^k(b) a_k^k$. Nous avons

$$e_k^* = n_k^k(\hat{\theta}) - x_k^T \beta^*$$

où

$$z_k = g_k^k(p(s)) A^{-1} e_k^*$$

avec β^* obtenu à partir de $\hat{\beta}$ en remplaçant y_k par $n_k^k(\hat{\theta})$. Notons que les résidus e_k^* ont la même forme que les résidus GREG e_k avec y_k remplacé par $n_k^k(\hat{\theta})$. En général, la solution $\hat{\theta}$ des équations d'estimation (3.10) peut ne pas être exprimable sous forme d'une fonction des totaux estimés. Par conséquent, nous suivons l'approche de Binder (1983) et écrivons l'estimateur par linéarisation de la matrice des covariances de $\hat{\theta}$ sous la forme

$$v^T(\hat{\theta}) = [f(\hat{\theta})]^{-1} \sum_s \sigma_s [f(\hat{\theta})]^{-1}, \quad (3.11)$$

où $f(\hat{\theta}) = -\partial S(\hat{\theta}) / \partial \hat{\theta}$ et $\sum_s \sigma_s(\hat{\theta})$ est la matrice des covariances estimée $v^T S(\hat{\theta}) v$ (3.11). En notant que $S(\hat{\theta})$ est un vecteur de totaux estimés dont les poids GREG sont $d_k^k(s) g_k^k(p(s))$, il découle de (3.6) et (3.11) que

$$v^T(\hat{\theta}) = v^T(z) \quad (3.12)$$

où

$$z_k = [f(\hat{\theta})]^{-1} g_k^k(p(s)) e_k^* \quad (3.13)$$

avec $e_k^* = (e_k^{k1}, \dots, e_k^{kp})^T$ et

$$e_k^{kj*} = n_k^k(\hat{\theta}) - x_k^k \beta^{kj*}, \quad j = 1, \dots, p.$$

En outre, nous obtenons $\hat{\beta}^{ju}$ à partir de $\hat{\beta}_j$ en remplaçant y_k par $n_k^k(\hat{\theta})$ et $v(z)$ est la matrice de covariance estimée du vecteur de totaux estimés $Z = \sum d_k^k(s) z_k$, où $n_k^k(\hat{\theta})$ est le j^e élément de $n_k^k(\hat{\theta})$. Le résultat (3.12) concorde avec l'estimateur de la variance par le jackknife linéarisé, voir pour l'échantillonnage stratifié à plusieurs degrés obtenu par Rao, Yung et Hidiroglou (2002).

Le résultat (3.12)-(3.13) peut aussi être obtenu directement en écrivant $\hat{\theta}$ sous la forme $f(p(s))$ et en évaluant $z_k = \partial f(p(s)) / \partial b_k |_{b=p(s)}$. Nous dénotons par $f(\hat{\theta})$ la solution de $\sum (b_k g_k^k(p(s)) n_k^k(\hat{\theta})) = 0$, c'est-à-dire

$$\sum (b_k g_k^k(p(s)) n_k^k(\hat{\theta}(b))) = 0, \quad (3.14)$$

avec les constantes spécifiées c_k^k et $\mathbf{X} = \sum d^k(s) x_k^k$ (consulter Särndal et coll. 1989). L'estimateur par le quotient, \hat{Y}^w , est un cas spécial pour lequel $q = 1$ (c'est-à-dire, x_k^k scalaire) et $c_k^k = x_k^k$, et $g^k(\mathbf{p}(s))$, donné par (3.1), se réduit à X/\hat{X} .

L'estimateur GREG peut être exprimé sous forme d'une fonction différentiable de totaux estimés. Donc, nous pouvons appliquer la théorie générale de la section 2 et il nous reste à évaluer $z_k^k = \partial f(\mathbf{b}) / \partial b^k |_{b=d(s)}$, où $f(\mathbf{b}) = \sum (b^k g^k(\mathbf{b}))^{y_k^k}$ est obtenu en remplaçant $\mathbf{p}(s)$ par \mathbf{b} dans la formule de \hat{Y}^w . En notant que $\partial A(\mathbf{b})^{-1} / \partial b^k = -A(\mathbf{b})^{-1} (\partial A(\mathbf{b}) / \partial b^k) A(\mathbf{b})^{-1}$, où $A(\mathbf{b}) = \sum b^k c_k^k x_k^k x_k^k$, nous obtenons

$$\partial(b^k g^k(\mathbf{b})) / \partial b^k = g^k(\mathbf{b}) - \mathbf{x}_T^T A(\mathbf{b})^{-1} b^k c_k^k x_k^k = g^k(\mathbf{b}) - \mathbf{x}_T^T A(\mathbf{b})^{-1} (b^k c_k^k x_k^k) \quad (3.2)$$

$$\begin{aligned} & - (X - \hat{X}) A(\mathbf{b})^{-1} (c_k^k x_k^k x_k^k) A(\mathbf{b})^{-1} (b^k c_k^k x_k^k) \cdot (3.3) \\ & = - \mathbf{x}_T^T A(\mathbf{b})^{-1} (b^k c_k^k x_k^k) \\ & \quad \partial(b^k g^k(\mathbf{b})) / \partial b^k \end{aligned}$$

Il découle maintenant de (3.2) et (3.3) que

$$\begin{aligned} & \partial f(\mathbf{b}) / \partial b^k = g^k(\mathbf{b}) e_k(\mathbf{b}), \quad (3.4) \\ & \text{où} \quad e_k(\mathbf{b}) = y_k^k - \mathbf{x}_T^T \mathbf{B}(\mathbf{b}) \quad (3.5) \end{aligned}$$

$$\begin{aligned} & \text{avec } \mathbf{B}(\mathbf{b}) = A^{-1}(\mathbf{b}) (\sum b^k c_k^k x_k^k y_k^k). \text{ Par conséquent, } z_k^k = \\ & \partial f(\mathbf{b}) / \partial b^k |_{b=d(s)} \text{ se réduit à} \quad (3.6) \end{aligned}$$

$$\text{où } e_k^k = y_k^k - \mathbf{x}_T^T \mathbf{B}(\mathbf{p}(s)) \text{ avec } \mathbf{B} = \mathbf{B}(\mathbf{p}(s)).$$

L'estimateur de la variance de \hat{Y}^w , résultant de (3.6), à savoir $v(z)$, tient compte des poids g_k^k , $g^k(\mathbf{p}(s))$, contrairement à l'estimateur de la variance par linéarisation type (voir, par exemple, Särndal et coll. 1991, page 237). Il concorde avec l'estimateur de la variance assisté par modèle de Särndal et coll. (1989). Il concorde aussi avec l'estimateur de la variance par le jackknife linéarisé lorsque ce dernier est applicable (Yung et Rao 1996).

3.2 Une classe générale de poids calés par régression

Nous examinons maintenant une classe générale de poids calés par régression de la forme $w_k^k(s) = d^k(s) h^k(\mathbf{p}(s))$

$$h^k(\mathbf{p}(s)) = 1 + (X - \hat{X})^T \tilde{\mathbf{Q}} |_{c_k^k} c_k^k x_k^k |, \quad (3.7)$$

Les poids calés $w_k^k(s)$ peuvent être réécrits sous la forme (Montanari 1998).

$$w_k^k(s) = d^k(s) + (X - \hat{X})^T \tilde{\mathbf{Q}} |_{c_k^k} c_k^k x_k^k | \quad (3.8)$$

$$\text{où } d^k(s) = d^k(s) d^k(s) / [E[d^k(s) d^k(s)]], \text{ et } c_k^k = c_k^k d^k(s) d^k(s) / [E[d^k(s) d^k(s)]] \text{ et}$$

$$\tilde{q}^{ab} = \sum_{k=1}^N d^k(s) c_k^k x_k^k x_k^k b^k + \sum_{k=1}^N d^k(s) c_k^k x_k^k x_k^k b^k.$$

Notons que $E[d^k(s)] = 1$ et $E[d^k(s)] = 1$. Si $d^k(s) = (1/\pi_k^k) a^k(s)$, alors $d^k(s)$ se réduit à $d^k(s) = a^k(s) a^k(s) / \pi_k^k$ et $c_k^k = (\pi_k^k - \pi_k^k) / (\pi_k^k \pi_k^k)$. Nous pouvons considérer l'estimateur par calage \hat{Y}^w résultant de (3.8) comme étant une fonction de totaux, en exprimant une forme quadratique sous forme d'un total de variables synthétiques (Sitter et Wu 2002). Par conséquent, nous pouvons utiliser la méthode de la section 2 et écrire $\hat{Y}^w = f(\mathbf{p}(s))$, où $\mathbf{p}(s) = \sum d^k(s) h^k(\mathbf{p}(s)) y_k^k$ où $\mathbf{p}(s)$ et $\mathbf{p}(s)$ est le vecteur d'éléments $d^k(s)$, $k < L$, arrangé en une série. Maintenant, de la même façon que nous avons calculé (2.3), nous obtenons

$$\hat{Y}^w - Y \approx \sum_{k < L} z_k^k (d^k(s) - 1) + 2 \sum_{k < L} z_k^k (d^k(s) - 1) \quad (3.9)$$

$$z_k^k = \partial f(\mathbf{p}(s)) / \partial b^k |_{b=d(s)} = 1, \quad z_k^k = \partial f(\mathbf{p}(s)) / \partial b^k |_{b=d(s)} = 1,$$

Puisque $v(z)$ représente par $v(z)$, nous remplaçons \hat{z}_k^k par $z_k^k = \partial f(\mathbf{p}(s)) / \partial b^k |_{b=d(s)}$ et \hat{z}_k^k par $z_k^k = \partial f(\mathbf{p}(s)) / \partial b^k |_{b=d(s)}$. Malheureusement, l'estimateur de la variance $v(z)$ pour obtenir $v(z)$.

$\hat{Y}^R = [(\sum d_k(s) y_k) / (\sum d_k(s) x_k)] X$, alors $m = 2$, $y^{1k} = y_k$, $y^{2k} = x_k$ et $f(\mathbf{1}, \mathbf{A}^y)$ se réduit au total X , en notant que $(Y/X)X = Y$. Notons que \hat{Y}^R est une fonction de $d(s)$, y et x et du total connu X , mais que nous laissons tomber X par souci de simplicité et que nous écrivons $\hat{Y}^R = f(d(s), y, x)$. La linéarisation de Taylor de $\hat{\theta}$ autour de Y donne l'approximation

$$\sqrt{n} N^{-a} (\hat{\theta} - \theta) \approx \frac{N}{\sqrt{n}} (\partial g(a) / \partial a)^T \big|_{a=Y} (Y - Y) \quad (2.1)$$

où $\partial g(a) / \partial a = (\partial g(a) / \partial a^1, \dots, \partial g(a) / \partial a^m)^T$ et $N^{-a} g(\cdot)$ tend vers une limite pour une certaine valeur de $a \geq 0$. La normalité asymptotique de $\sqrt{n} N^{-a} (\hat{\theta} - \theta)$ découle de (2.1), à condition qu'un théorème central limite pour $\sqrt{n} N^{-1} (Y - Y)$ soit vérifié et que $g(\cdot)$ ait des dérivées premières continues au voisinage de la moyenne \bar{Y} . Krewski et Rao (1981) ont justifié (2.1) pour l'échantillonnage stratifié.

Soit $\hat{Y} = \sum b_j Y^j$ pour des nombres réels arbitraires $b = (b^1, \dots, b^N)^T$, et $g(\hat{Y}) = f(\hat{\theta}, \mathbf{A}^y) = f(\hat{\theta})$. En notant que $\hat{Y} = \mathbf{A}^y d(s)$ et $Y = \mathbf{A}^y \mathbf{1}$, nous pouvons exprimer (2.1) sous la forme

$$\sqrt{n} N^{-a} (\hat{\theta} - \theta) \approx \frac{N}{\sqrt{n}} (\partial g(\hat{Y}) / \partial \hat{Y})^T \big|_{\hat{Y}=\mathbf{A}^y} \mathbf{A}^y (d(s) - \mathbf{1})$$

$$= \frac{N}{\sqrt{n}} \sum_{k=1}^K (\partial f(b) / \partial x_k^T) b^{k-1} Y^k (d_k(s) - 1) \cdot (2.2)$$

en notant que $\hat{X} = Y$ est équivalent à $b = \mathbf{1}$. Maintenant, nous substituons $Y^k = \partial X / \partial b^k \big|_{b=\mathbf{1}}$ dans (2.2) pour obtenir

$$\sqrt{n} N^{-a} (\hat{\theta} - \theta) \approx \frac{N}{\sqrt{n}} \sum_{k=1}^K (\partial f(b) / \partial b^k) \big|_{b=\mathbf{1}} (d_k(s) - 1) \quad (2.3)$$

où $\hat{z} = (z^1, \dots, z^K)^T$ avec $z_k = \partial f(b) / \partial b^k \big|_{b=\mathbf{1}}$. Un estimateur de la variance du deuxième membre de (2.3) est donné par $(n/N^2) v(\hat{z})$, où $v(\hat{z})$ est l'estimateur de la variance du total estimé $\sum d_k(s) z_k$ par $z_k = z_k^*$ sont inconnues, nous remplaçons z_k^* par $z_k = \partial f(b) / \partial b^k \big|_{b=p(s)}$, pour obtenir $(n/N^2) v(z)$. Donc, un estimateur de la variance par linéarisation de $\hat{\theta}$ est donné par

$$v^{(7)}(\hat{\theta}) = (N^2 n / N^2) v(z), \quad (2.4)$$

qui se réduit à $v(z)$ si $\alpha = 1$. Notons que $v^{(7)}(\hat{\theta})$ donné par (2.4) s'obtient simplement à partir de la formule $v(y)$ pour Y en remplaçant y_k par z_k pour $k \in s$. Notons que nous ne commençons pas par calculer les dérivées partielles $\partial f(b) / \partial b^k$ à $b = \mathbf{1}$ pour obtenir \hat{z} , pour ensuite substituer les estimations aux composantes inconnues de \hat{z} . Par conséquent, l'esprit de notre méthode est comparable à

3. ESTIMATEURS PAR CALAGE

L'estimateur par le quotient peut être considéré comme un estimateur par calage, $\hat{Y}^R = \sum w_k(s) Y^k$, dont les poids $w_k(s) = (X/\bar{X}) d_k^R(s)$ sont explicites et obéissent à la contrainte de calage $\sum w_k(s) x_k = X$. Les estimateurs par calage d'un total Y de la forme $\hat{Y}^w = \sum w_k(s) Y^k$ dont les poids $w_k(s)$ sont explicites et satisfont aux contraintes de calage $\sum w_k(s) x_k = X$ sont utilisés à grande échelle, où $\mathbf{x}_k = (x_{1k}, \dots, x_{qk})^T$ et $\mathbf{X} = (X^1, \dots, X^q)^T$ est le vecteur de totaux connus de variables auxiliaires x_j , $j = 1, \dots, q$. À la sous-section 3.1, nous considérons l'estimateur par la régression généralisée (GREG), puis, à la sous-section 3.2, nous étudions une classe générale d'estimateurs calés par régression. L'extension aux estimateurs, $\hat{\theta}$, obtenus comme solutions des équations d'estimation est présentée à la sous-section 3.3. Le cas des estimateurs par calage généraux est examiné à la sous-section 3.4.

L'approche de Binder. L'estimateur de la variance $v^{(7)}(\hat{\theta})$ est valide parce que z_k^* est un estimateur convergent de z_k^* .

Exemple 2.1 Supposons que $\hat{\theta}$ est l'estimateur par le quotient $\hat{Y}^R = X [(\sum d_k(s) y_k) / (\sum d_k(s) x_k)]$ du total Y . Alors $f(\hat{\theta}) = X [(\sum b^k x_k y_k) / (\sum b^k x_k)]$ et

$$\partial f(b) / \partial b^k = \frac{X \sum b^k x_k y_k}{\sum b^k x_k} - \hat{Y}^R x_k = \frac{X}{\sum b^k x_k} (y_k - \hat{Y}^R x_k)$$

Par conséquent,

$$z_k = \partial f(b) / \partial b^k \big|_{b=p(s)} = \frac{X}{\sum b^k x_k} (y_k - \hat{Y}^R x_k)$$

Et par conséquent,

découle de (1.11), qu'un estimateur de la variance par linéarisation de $T(F)$ est donné par

$$v_L[T(F)] = v[(z - R)/N], \quad (1.12)$$

où

$$z_k = IC(y_k, F, T), \quad (1.13)$$

et

$$R = \sum_{k \in s} d^k(s) z_k / \sum_{k \in s} d^k(s). \quad (1.14)$$

Pour éviter le calcul des z_k , Campbell (1980) propose un estimateur par le jackknife de z_k pour chaque $k \in s$. Il est donné par

$$z_{kL} = \frac{1 - \bar{d}^k(s)}{T(F) - T(F_{-k})}, \quad (1.15)$$

où

$$dF_{-k}^k(y) = \begin{cases} \frac{dF(y) - \bar{d}^k(s)}{1 - \bar{d}^k(s)} & \text{si } y = y_k \\ \frac{dF(y)}{1 - \bar{d}^k(s)} & \text{si } y \neq y_k. \end{cases} \quad (1.16)$$

L'estimateur de la variance par linéarisation résultant est donné par $v[(z_L - R)/N]$. Notons que la méthode du jackknife proposée diffère du jackknife habituel pour l'échantillonnage, pour l'échantillonnage stratifié à plusieurs phases, la méthode habituelle du jackknife comprend l'élimination à tour de rôle de grappes d'échantillonnage, tandis que la méthode de Campbell comprend l'élimination à tour de rôle d'unités d'échantillonnage. En outre, le jackknife habituel n'est pas toujours applicable (par exemple, échantillonnage avec probabilités inégales sans remise) contrairement à la méthode de Campbell qui utilise l'estimateur de la variance sans biais $v(y)$ du total Y pour le plan de sondage donné, puis remplace Y par $(z_L - R)/N$. Cependant, les calculs que nécessite la méthode de Campbell peuvent être très lourds, parce qu'il faut calculer $T(F_{-k})$ pour chaque unité $k \in s$; dans le cas des enquêtes à grande échelle, le nombre d'unités d'échantillonnage peut être très grand, comme pour l'Enquête sur la population active du Canada.

Deville (1999) et Berger (2002) obtiennent des résultats fort comparables à ceux de Campbell (1980). Au lieu d'utiliser la mesure de probabilité naturelle F , ils considèrent des fonctionnelles de la forme $T(M)$, où M représente une mesure qui attribue le poids de sondage $d^k(s)$ à tout point y_k pour k dans s et une valeur nulle aux unités k qui ne sont pas comprises dans s . Par exemple, $T(M) = \int x dM(x) = \sum d^k(s) y_k$ si le paramètre de population est le total

suivante :

$$\sqrt{n} N^{-\alpha} [T(M) - T(M)] \sum \frac{N}{\sqrt{n}} (d^k(s) - 1) z_k, \quad (1.17)$$

où $d^k(s) = 0$ si k n'est pas compris dans l'échantillon s . En outre $z_k = IT(M; y_k)$ avec IT représentant la fonction d'influence de $T(M)$ définie par

$$IT(M; y) = \lim_{t \rightarrow 0} \frac{1}{t} [T(M + t\delta_y) - T(M)]. \quad (1.18)$$

Comme nous l'avons mentionné plus haut, il n'est pas facile de justifier l'approximation (1.17) pour des fonctionnelles générales $T(\cdot)$. Deville (1999) a élaboré des règles pour évaluer $IT(M; y)$ pour certaines fonctionnelles $T(M)$. Berger (2002) utilise la méthode du jackknife pour estimer $z_k = IT(M; y_k)$, de la même façon que Campbell (1980).

En notant que $\sum d^k(s) z_k = Y(\bar{z})$, il découle de (1.17) qu'un estimateur de la variance par linéarisation de $N^{-\alpha} T(M)$ est donné par $N^{-2} v(\bar{z})$. Mais z_k dépend de paramètres inconnus et l'estimateur correspondant, z_k , pourrait ne pas être unique. Par exemple, supposons que $T(M) = Y^R = (Y/X)X$, alors $\alpha = 1$ et $\bar{z}_k = Y_k - R X_k$, où dans ce cas, deux candidats possibles pour z_k sont $z_k = Y_k - R X_k$ et $z_k = (X/X)(Y_k - R X_k)$. Donc, le choix de z_k en présence d'information auxiliaire, comme un total connu X , n'est pas unique sous l'approche de Deville. Contrairement à cette dernière, notre méthode ne produit qu'un seul choix z_k et permet d'éviter le calcul de z_k pour déterminer z_k . Notre z_k possède les propriétés désirables mentionnées à la section 1, du moins dans un certain nombre de cas importants.

2. LA MÉTHODE

Pour justifier la méthode, nous commençons par un cas général simple où l'estimateur $\hat{\theta}$ d'un paramètre θ peut être exprimé comme une fonction lisse $g(X)$ de totaux estimés $Y = (Y_1, \dots, Y_t, \dots, Y_m)^T$, où $Y_t = \sum_{k \in U} d^k(s) y_k$, $t = 1, \dots, m$, est un estimateur du total $Y_t = \sum_{k \in U} Y_t^{(k)}$, et $\theta = g(X)$ avec $X = (X_1, \dots, X_t, \dots, X_m)^T$. Nous pouvons écrire θ sous la forme $\theta = f(d(s), A_y)$ et $\theta = f(1, A_y)$, où A_y est une matrice de dimensions $m \times N$ dont la k^e colonne est $Y_k = (y_k, \dots, y_k, \dots, y_k)^T$, $k = 1, \dots, N$, $d(s) = (d(s), \dots, d(s))^T$ et 1 est le vecteur de dimension N de valeur 1. Par exemple, si $\hat{\theta}$ représente l'estimateur par le quotient

exemple, si $T(F)$ est la moyenne de population $\int y dF(y)$, alors $T(F) = \int y dF(y) = n^{-1} \sum_{k=1}^n y_k = \bar{y}$, la moyenne d'échantillon. Notons que F attribue une masse égale, $1/n$, à chaque valeur d'échantillon y_1, \dots, y_n . Si T est « suffisamment régulière », alors $T(F)$ peut être linéarisée près de F en ce qui concerne la courbe (ou fonction) d'influence de $T(\cdot)$ donnée par

$$IC(y; F, T) = \lim_{a \rightarrow 0} [T((1-a)F + a\delta_y) - T(F)] / a, \quad (1.5)$$

où δ_y représente la mesure de masse 1 portée par y . Nous

avons

$$\sqrt{n}[T(\hat{F}) - T(F)] = \sqrt{n} \int IC(y; F, T) d\hat{F}(y) + \sqrt{n} R_n \quad (1.6)$$

$$= \frac{1}{n} \sum_{k=1}^n \hat{z}_k + \sqrt{n} R_n$$

où $\hat{z}_k = IC(y_k; F, T)$ et $\sqrt{n} R_n$ est un terme de reste. Si $\sqrt{n} R_n$ converge en probabilité vers zéro quand $n \rightarrow \infty$ (représenté par $\sqrt{n} R_n \xrightarrow{p} 0$), il découle alors de (1.6) que $\sqrt{n}[T(\hat{F}) - T(F)]$ est asymptotiquement normal de moyenne 0 et de variance

$$A(F, T) = \int [IC(y; F, T)]^2 dF(y), \quad (1.7)$$

en notant que les termes \hat{z}_k dans (1.6) sont des variables aléatoires iid. Comme l'a souligné Huber (1981, page 13), $\sqrt{n} R_n$ est « souvent » asymptotiquement négligeable, mais la preuve de cette propriété n'est pas nécessairement facile pour des fonctionnelles générales $T(F)$. Serfling (1980, section 6.2) donne les deux conditions suivantes pour $\sqrt{n} R_n \xrightarrow{p} 0$, applicables aux variables aléatoires générales y_1, \dots, y_n (pas nécessairement iid) : i) $T(\cdot)$ est « stochastiquement dérivable » en F ; ii) $\sqrt{n} \sup |F(y) - F(y)|$ est bornée en probabilité, où \sup est sur y . La condition (ii) est satisfaite dans le cas iid, mais il n'est pas toujours facile de prouver (ii) pour des plans d'échantillonnage complexes. La condition (i) signifie qu'il existe une fonctionnelle $T(F; F_n - F)$ telle que $T(F_n) - T(F) = n^{-1} \sum_{k=1}^n (F(y_k) - F(y))$ quand cette dernière tend vers zéro. Cette condition pourrait ne pas être facile à vérifier qu'en pratique, il est plus efficace d'analyser directement R_n par la méthode des inéquations différentielles ».

Un estimateur naturel de la variance asymptotique

$$\frac{1}{n(n-1)} \sum_{k=1}^n \sum_{\ell \neq k}^n \hat{z}_k \hat{z}_\ell = \frac{n-1}{n} \sum_{k=1}^n [T(\hat{F}^{-k}) - T(\hat{F})]^2; \quad (1.10)$$

où $IC(y; F, T)$ est la courbe d'influence évaluée en $F = F$. Il s'ensuit qu'un estimateur de la variance par linéarisation de $T(F)$ est

$$v_L[T(\hat{F})] = A(F, T)/n. \quad (1.9)$$

La mise en oeuvre pratique de $v_L[T(\hat{F})]$ comprend le calcul de $IC(y_k; F, T)$ pour chaque T spécifiée. Ce calcul peut être évité en utilisant la méthode du jackknife. En substituant \hat{F} à F et $1/(n-1)$ à a dans (1.5), nous obtenons un estimateur par le jackknife de $IC(y_k; F, T)$ sous la forme $\hat{z}_{kj} = (n-1)[T(\hat{F}) - T(\hat{F}^{-k})]$, où \hat{F}^{-k} est la fonction de distribution empirique obtenue quand on omet y_k . L'estimateur de la variance par le jackknife résultant $T(\hat{F})$ est

Campbell (1980) a essayé d'étendre les résultats sus-mentionnés obtenus pour le cas iid aux plans d'échantillonnage généraux, en utilisant les poids de sondage $d_k(s)$. Le paramètre de population (ou de recensement) θ est maintenant donné par $\theta = T(F^N)$, où F^N est la fonction de distribution de population qui attribue une masse égale, $1/N$, à chacune des N valeurs de population y_1, \dots, y_N . Une fonction de distribution empirique est donnée par $\hat{F}(y) = \sum_{k \in s} d_k(s) I(y_k \leq y)$, où $d_k(s) = d_k(s) / \sum_{k \in s} d_k(s)$ sont les poids de sondage normalisés. Notons que $\hat{F}(y)$ attribue la masse $\hat{d}_k(s)$ à l'élément $k \in s$. Un estimateur de $\theta = T(F^N)$ est donné par $\hat{\theta} = T(\hat{F})$. Par exemple, si $T(F^N)$ est la moyenne de population $\int y dF^N(y)$, alors $T(\hat{F}) = \int y d\hat{F}(y) = \sum_{k \in s} \hat{d}_k(s) y_k$. $T(\hat{F})$ est asymptotiquement normal de moyenne 0 et de variance

$$A(F^N, T) = n \text{Var} \left[\sum_{k \in s} d_k(s) \hat{z}_k / \sum_{k \in s} d_k(s) \right]$$

$$\approx n \text{Var} \left[\sum_{k \in s} d_k(s) \{(\hat{z}_k - R)/N\} \right], \quad (1.11)$$

en utilisant la variance approximative d'un ratio, où $R = \sum_{k \in s} \hat{z}_k / N$ est la moyenne de population des \hat{z}_k et $\hat{z}_k = IC(y_k; F^N, T)$. Si nous représentons l'estimateur sans biais de la variance de $\hat{Y} = \hat{Y}(y) = \sum_{k \in s} d_k(s) y_k$ par $v(y)$, il

comme si les unités primaires d'échantillonnage étaient sélectionnées avec remise. Notons que la méthode du jackknife n'est pas applicable de façon générale à tout plan d'échantillonnage.

Pour l'estimateur $\hat{\theta} = g(X_1^m, \dots, X^m)$ d'une fonction lisse de totaux, $\theta = g(X_1, \dots, X^m)$, la méthode de Binder (1996) donne

$$\hat{\theta} - \theta = \sum d^k(s) z_k + \dots$$

avec

$$z_k = \sum_{m=1}^I \partial g(\mathbf{a}) / \partial a_i \Big|_{\mathbf{a}=\mathbf{y}} y_{ki}, \quad (1.4)$$

où $\mathbf{Y} = (Y_1^m, \dots, Y^m)^T$ et $\mathbf{a} = (a_1, \dots, a_m)^T$. Il découle de (1.4) que les dérivées partielles $\partial g(\mathbf{a}) / \partial a_i$ sont évaluées à \mathbf{Y} pour obtenir les z_k , tandis que dans le cas de la méthode standard (voir, par exemple, Andersson et Nordberg 1994), on les évalue à $\mathbf{Y} = (Y_1^m, \dots, Y^m)^T$ avant d'obtenir z_k , puis on substitue les estimations aux composantes inconnues. Par exemple, pour l'estimateur par le quotient Y^R , le terme X/Y disparaît de z_k dans la méthode standard, parce que X/Y devient 1 quand on remplace X par X .

Bien que la méthode de Binder (1996) soit simple et séduisante, une méthode applicable plus rigoureusement et généralement est nécessaire. À la section 2, nous proposons une autre méthode que celle justifiable théoriquement et qui, par la même occasion, mène directement à un estimateur de la variance de type Y^R .

À la section 3, nous appliquons la méthode à divers problèmes, dont les estimateurs par régression-calage d'un total Y et d'autres estimateurs définis explicitement ou implicitement comme étant des solutions d'équations de estimation, par exemple les estimateurs des paramètres de régression logistique avec poids de sondage calés sur des totaux de population auxiliaires connus. Nous obtenons aussi un nouvel estimateur de la variance pour une classe générale d'estimateurs par calage qui inclut l'estimateur par la méthode itérative du quotient (raking ratio) généralisée et des estimateurs par régression généralisée. À la section 4, nous étendons la méthode proposée à l'échantillonnage à deux phases pour obtenir un estimateur de la variance qui utilise plus complètement les données de l'échantillon de première phase que les estimateurs de la variance par linéarisation classiques.

Pour le cas de variables aléatoires indépendantes et identiquement distribuées (iid) y_1^m, \dots, y^m dont la fonction de distribution est $F(y)$, l'étude de l'estimation de paramètres généraux $\theta = T(F)$ est décrite par de nombreux auteurs (voir, par exemple, Huber 1981). Un estimateur naturel de $\theta = T(F)$ est $\hat{\theta} = T(\hat{F})$, où $\hat{F}(y)$ est la fonction de distribution empirique donnée par $\hat{F}(y) = n^{-1} \sum_{k=1}^n I(y_k \leq y) = I(y_k \leq y) = 1$ si $y_k \leq y$ et $I(y_k \leq y) = 0$ si $y_k > y$. Par

Bindert (1996) présente une approche « livre de recettes » élégante de la linéarisation de Taylor qui mène directement à des estimateurs de la variance par linéarisation de type Y^R . Il applique la méthode à des fonctions lisses de totaux estimés, $g(X_1^m, \dots, X^m)$, à des estimateurs par régression généralisée et à la statistique de la somme des rangs de Wilcoxon (Wilcoxon rank sum statistic). Pour illustrer la méthode de Binder, considérons un estimateur par le quotient

$$Y^R = (Y/X)X = RX,$$

où $Y = \sum_{k=1}^N d^k(s) y_k = Y(y)$, $X = \sum_{k=1}^N d^k(s) x_k = X(x)$ et les $d^k(s)$ sont les poids de sondage avec $d^k(s) = 0$ si l'élément de population k n'est pas dans l'échantillon, par exemple $d^k(s) = (1/\pi_k)/a^k(s)$ où π_k est la probabilité d'inclusion de l'élément k dans l'échantillon, $a^k(s) = 1$ si $k \in s$, $a^k(s) = 0$ autrement et \sum représente la sommation sur les éléments de la population. Nous supposons que les poids produisent un estimateur sans biais par rapport au plan de sondage Y de Y , c'est-à-dire $E(d^k(s)) = 1$ pour $k = 1, \dots, N$. Maintenant, prenons la dérivée totale de Y^R pour obtenir

$$(dY^R)_X = (dR)_X = \frac{dX}{X} [dY - R(dX)], \quad (1.1)$$

et remplaçons toutes les dérivées totales figurant dans (1.1) par les écarts des estimateurs par rapport à leurs paramètres de population respectifs, par exemple remplaçons dY^R par $Y^R - Y$. Alors (1.1) donne

$$Y^R - Y = \sum d^k(s) z_k - \frac{X}{Y} (Y - RX), \quad (1.2)$$

où

$$z_k = \frac{X}{Y} (y_k - R x_k). \quad (1.3)$$

Dans (1.2), le terme $\sum d^k(s) z_k$ se réduit à zéro, mais nous le retenons pour l'estimation de la variance. Par contre, nous ne tenons pas compte du dernier terme de (1.2) dans cette estimation. Donc, $Y^R - Y$ est représenté par $\sum d^k(s) z_k = Y(2)$ aux fins de l'estimation de la variance. Si nous représentons un estimateur sans biais de la variance de $Y^R = Y(y)$ par $v(y)$, l'estimateur de la variance de Y^R de Binder est donné par $v(2)$. L'estimateur de la variance par linéarisation $v(2)$, obtenu à partir de (1.3), concorde avec v^R pour l'échantillonnage aléatoire simple et l'échantillonnage stratifié à plusieurs phases si nous traitons l'échantillon

Estimateurs de variance par linéarisation pour des données d'enquête

ARDELLATIF DEMNATI et J.N.K. RAO¹

RÉSUMÉ

En échantillonnage, on utilise souvent la linéarisation de Taylor pour obtenir des estimateurs de variance pour des données de recensement, comme des ratios, ou des coefficients de régression et de corrélation, qui peuvent être exprimés sous forme de fonctions linéaires de totaux. La linéarisation de Taylor est généralement applicable à tout plan d'échantillonnage, mais elle peut produire de multiples estimateurs de variance qui sont asymptotiquement sans biais par rapport au plan de sondage sous échantillonnage répété. Pour choisir lequel de ces estimateurs utiliser, il faut tenir compte d'autres critères, comme i) l'absence approximative de biais pour la variance par rapport au modèle de l'estimateur sous un modèle hypothétique, et ii) la validité sous l'échantillonnage répété conditionnel. Dans le présent article, nous proposons une nouvelle approche pour calculer les estimateurs de variance par linéarisation de Taylor. Elle mène directement à un estimateur de variance qui satisfait aux critères susmentionnés, du moins dans un nombre important de cas. Nous appliquons la méthode à divers problèmes, qui englobent les estimateurs d'un total, ainsi que d'autres estimateurs définis explicitement ou implicitement comme solutions d'équations d'estimation. En particulier, nous étudions les estimateurs des paramètres de régression logistique avec poids calés. Cette étude nous mène à un nouvel estimateur de la variance pour une classe générale d'estimateurs par calage qui inclut l'estimateur par la méthode itérative du quotient (ratio) généralisée et les estimateurs par régression qui utilisent plus complètement les données de l'échantillon de première phase que les estimateurs de variance par linéarisation classiques.

MOTS CLÉS : Calage; poids de sondage; équations d'estimation; estimateur par la méthode itérative du quotient (ratio) (ratio); estimateurs par régression; échantillonnage à deux phases.

1. INTRODUCTION

La linéarisation Taylor est une méthode très répandue pour estimer la variance de la variance pour des statistiques complexes, comme les estimateurs des coefficients de régression, ainsi que les estimateurs par rapport au plan d'échantillonnage et de la population. Elle s'applique généralement à tout plan d'échantillonnage qui permet une estimation sans biais de la variance des estimateurs linéaires et demande des calculs plus simples qu'une méthode de rééchantillonnage, comme le jackknife. Cependant, elle peut produire des estimateurs multiples de la variance qui sont asymptotiquement sans biais par rapport au plan de sondage sous échantillonnage répété. Par conséquent, pour déterminer lequel de ces estimateurs il convient d'utiliser, il faut tenir compte d'autres critères comme i) l'absence approximative de biais pour la variance de l'estimateur par rapport au modèle sous un modèle hypothétique et ii) la validité sous l'échantillonnage conditionnel répété. Par exemple, dans le contexte de l'échantillonnage aléatoire simple et de l'estimateur par le quotient, $\bar{y}_r = (\bar{y}/x) \bar{X}$, du total de population \bar{Y} , Royall et Cumberland (1981) montrent qu'un estimateur de la variance par linéarisation utilisé couramment, $v_L = N^2(n^{-1} - N^{-1})s_{\bar{y}}^2$, ne capte pas la variance conditionnelle de \bar{Y}_r sachant \bar{x} , contrairement à l'estimateur de

la variance par le jackknife v_J . Ici, \bar{y} et \bar{x} sont les moyennes d'échantillon, \bar{X} est le total connu de population d'une variable auxiliaire x , $s_{\bar{y}}^2$ est la variance d'échantillon des résidus $z_k = y_k - (\bar{y}/\bar{x})x_k$ et (n, N) représente les tailles d'échantillon et de population. Par linéarisation de l'estimateur jackknife de la variance, v_J , on obtient un estimateur de la variance par linéarisation différent, $v_{JL} = (X/\bar{x})^2 v_L$. Ce dernier capte la variance conditionnelle, ainsi que la variance inconditionnelle, où $\bar{X} = X/N$ est la moyenne de x . Par conséquent, on pourrait préférer utiliser v_{JL} ou v_J plutôt que v_L . Yung et Rao (1996) considèrent des estimateurs poststratifiés ajustés par régression généralisée et par la méthode du quotient sous échantillonnage stratifié à plusieurs phases et obtiennent un estimateur de la variance par le jackknife linéarisé, v_{JL} , en linéarisant v_J . Valliant (1993) obtient aussi v_{JL} pour l'estimateur poststratifié ajusté par la méthode du quotient et réalise une étude en simulation pour démontrer que v_J et v_{JL} possèdent tous deux de bonnes propriétés conditionnelles sachant les échantillonnages estimés dans les strates à posteriori. Särndal, Swensson et Wretman (1989) montrent que v_{JL} est à la fois asymptotiquement sans biais par rapport au plan de sondage et approximativement sans biais par rapport au modèle au sens de $E_m(v_{JL}) \approx V_m(\bar{Y}_r)$, où E_m représente l'espérance fondée sur le modèle et $V_m(\bar{Y}_r)$ représente la

¹ A. Demnati, Division des méthodes et statistiques sociales, Statistique Canada, Immeuble R.-H.-Coats, 15^e étage, Ottawa, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, (Ontario), Canada, K1S 5B6.

- BREWER, W.E. (1994). Autobiographical memory and survey research. Dans *Autobiographical Memory and the Validity of Retrospective Reports*, (Eds. S. Schwarz et S. Sudman). New York: Springer-Verlag.
- CANNELL, C.F., MILLER, P. et OKSENBURG, L. (1981). Research on interviewing techniques. Dans *Sociological Methodology 1981*, (Ed. S. Leinhardt). San Francisco: Jossey-Bass.
- EAGLY, A.H., et CHAIKEN, S. (1993). *The Psychology of Attitudes*. Orlando, FL: Harcourt, Brace, Jovanovich.
- GREENBERG, B.G., ABUL-ELA, A.T., SIMMONS W.R. et HORVITZ, D.G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.
- GRICE, H.P. (1975). Logic and conversation. Dans *Syntax and Semantics 3: Speech Acts*, (Eds. P. Cole et J.L. Morgan). New York: Academic Press. 41-58.
- HIPPLER, H.J., SCHWARZ, N. et SUDMAN, S. (Eds.) (1985). *Social Information Processing And Survey Methodology*. New York: Springer-Verlag.
- HORVITZ, D.G., SHAH, B.V. et SIMMONS, W.R. (1967). The unrelated question randomized response model. Dans *Proceedings of the Social Statistics Section, American Statistical Association*, 65-72.
- HUTTENLOCHER, J., HEDDES, L.V. et BRADBURN, N.M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology, Learning, Memory and Cognition*, 16, 196-213.
- JABINE, T., STRAF, M., TANUR, J. et TOURANGEAU, R. (Eds.) (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, D.C.: National Academy Press.
- KNAEUPER, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly*, 63, 347-370.
- KROSNIK, J.A., et ALWIN, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Research Quarterly*, 51, 201-219.
- LITTLE, R., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- LOFTUS, E.F., et MARRUBERGER, W. (1985). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*, 11, 114-120.
- MOSS, L., et GOLDSTEIN, H. (Ed.) (1979). Recall Method in Social Surveys. London, NFER Publishing Co., Ltd.
- NETER, J., et WAKSBERG, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.
- OMURCHEARAIGH, C., et MOUSTAKI, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, A*, 162, 2, 177-194.
- PETTY, R.E., et CACIOPPO, J.T. (1986). *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.
- WENTLAND, E.J., et SMITH, K.W. (1993). *Survey Responses: An Evaluation of Their Validity*. San Diego: Academic Press.
- SCHUMAN, H., et PRESSER, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic Press.
- SCHWARZ, N., et BLESS, H. (1992). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. Dans *The Construction of Social Judgments*, (Eds. L.L. Martin et A. Tesser). Hillsdale, N.J.: Erlbaum, 217-245.
- SCHWARZ, N., et KNAEUPER, B. (2000). Cognition, aging, and self-reports. Dans *Cognitive Aging: A Primer*, (Eds. D.C. Park et N. Schwarz). Philadelphia: Psychology, 233-252.
- SCHWARZ, N., MUENKEL, T. et HIPPLER, H.J. (1990). What determines a perspective? Contrast effects as a function of the dimension tapped by preceding questions. European *Journal of Social Psychology*, 20, 357-361.
- SCHWARZ, N., STRACK, F. et MAI, H.F. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, 55, 3-23.
- SCHWARZ, N., et SUDMAN, S. (Eds.) (1992). *Context Effect in Social and Psychological Research*. New York: Springer-Verlag.
- SCHWARZ, N., et SUDMAN, S. (Eds.) (1994). *Autobiographical Memory and the Validity of Retrospective Reports*. New York: Springer-Verlag.
- SUDMAN, S., et BRADBURN, N.M. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chichester: Aldine.
- SUDMAN, S., BRADBURN, N.M. et SCHWARZ, N. (1996). *Thinking About Answers*. San Francisco: Jossey-Bass.
- TANUR, J.M. (Ed.) (1992). *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys*. New York: Russell Sage Foundation.
- TOURANGEAU, R. (1984). Cognitive sciences and survey methods. Dans *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur et R. Tourangeau). Washington, D.C.: National Academy Press
- TOURANGEAU, R., et RASINSKI, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299-314.
- TOURANGEAU, R., RIPS, L.J. et RASINSKI, K. (2000). *The Psychology of Survey Response*. Cambridge, England: Cambridge University Press.
- TURNER, C.F., et MARTIN, E. (1982). *Surveys of Subjective Phenomena*. Cambridge, MA: Harvard University Press.
- WARNER, S.L. (1965). Randomized response: A survey technique for eliminating error answer bias. *Journal of the American Statistical Association*, 60, 63-69.

signification est communiquée entre les êtres humains. Les questions ont une signification que nous nous attendons à être comprise par les répondants. Il y a une limite à l'amélioration possible du processus de communication intelligible sans qu'on possède une compréhension beaucoup plus approfondie des mécanismes de base de la communication. Un effort multidisciplinaire concerté des linguistes, des psychologues, des statisticiens, des spécialistes de la cognition et d'autres sera nécessaire pour déchiffrer le code de la signification, au même titre que les spécialistes des sciences naturelles ont déchiffré le code génétique. Il s'agit là d'un des grands défis scientifiques de notre époque.

BIBLIOGRAPHIE

- BADDELEY, A. (1979). The limitations of human memory: Implications for the design of retrospective surveys. Dans *The Recall Method in Social Surveys*, (Eds. L. Moss et H. Goldstein). London: NFER Publishing Co., Ltd.
- BARSAULOU, L.W. (1988). The content and organization of autobiographical memories. Dans *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*, (Eds. U. Neisser et E. Winograd). Cambridge, England: Cambridge University Press.
- BELSON, W.A. (1968). Respondent understand of survey questions. *Polls*, 3(1), 1-13.
- BELSON, W.A. (1981). *The Design and Understanding of Survey Questions*. Aldershot, England: Cower.
- BIDDERMAN, A. (1980). *Report of a Workshop on Applying Cognitive Psychology to Recall Problems of the National Crime Survey*. Washington, D.C.: Bureau of Social Science Research.
- BINGHAM, W.V.D., et MOORE, B.V. (1934). *How to Interview*, (Ed. Révisée). New York: Harper Collins.
- BISHOP, G.F., OLSENDICK, R.W. et TUCHFARBBER, R.J. (1986). Opinions on fictitious issues: The precurs to answer survey questions. *Public Opinion Quarterly*, 50, 240-250.
- BRADBURN, N.M. (1992). What have we learned? Dans *Context Effects in Social and Psychological Research*, (Eds. N. Schwarz et S. Sudman). New York: Springer-Verlag.
- BRADBURN, N.M., et DANIS, C. (1984). Potential contributions of cognitive research to survey questionnaire design. Dans *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur et R. Tourangeau). Washington, D.C.: National Academy Press.
- BRADBURN, N.M., et SUDMAN, S. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.
- BRADBURN, N.M., SUDMAN, S. et ASSOCIATES (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.
- BREWER, W.E. (1986). What is autobiographical memory? Dans *Autobiographical Memory*, (Ed. D.C. Rubin). Cambridge, England: Cambridge University Press.

L'application de nos connaissances des processus cognitifs fondamentaux à l'étude de la formulation et de l'ordre des questions nous permet d'améliorer considérablement notre compréhension des effets de contexte. Grâce à la science de la cognition, nous progressons à grands pas vers la compréhension du fonctionnement du cerveau et de la façon dont nous organisons et traitons l'information. Le champ des connaissances dans ces domaines s'accroît rapidement. À mesure que nous en apprenons davantage, nombre des représentations conceptuelles décrites dans le présent essai évolueront et s'avéreront soit fausses soit très élaborées.

Enfin, de grand défis se posent dans le domaine de la linguistique. Nombre des effets dont nous avons discutés dans l'essai se manifestent à cause des ambiguïtés du langage. Comprendre comment la signification est encodée dans le langage et comment nous extrayons cette signification du langage parlé et écrit pose un défi de taille. Notre capacité à résoudre certains des problèmes les plus fondamentaux que pose la création des questionnaires dépendra peut-être plus que de toute autre chose des progrès réalisés dans ces domaines.

Quels sont les domaines fortement prioritaires en ce qui concerne la recherche? Dans le court terme, il faut, selon moi, s'efforcer de mieux comprendre les effets biaisés de la participation en baisse des répondants, particulièrement ceux des dispositions possibles des réponses produites par les répondants réticents. Nous devons développer des modèles d'effet de réponse qui tiennent compte non seulement des données manquantes, qu'il s'agisse de non-réponse partielle ou totale, mais aussi des effets de réponse introduits par les répondants réticents qui ne donnent que des réponses partielles ou des réponses qui ne sont pas bien réfléchies. Les modèles d'imputations multiples, tels que ceux élaborés par Little et Rubin (1987) et les approches basées sur les variables latentes, telles que développées par O'Muircheartaigh et Moustaki (1999) sont prometteurs. D'autres travaux empiriques sont nécessaires sur les effets des efforts en vue de pousser à répondre des personnes qui, au départ, ne désirent pas participer à une enquête.

Dans le long terme, il faudra poursuivre l'étude des mécanismes par lesquels les catégories de questions et de réponses suscitent l'élaboration cognitive et activent les pensées qui sont alors utilisées pour répondre aux questions. Nous devons savoir ce qui, dans certaines questions, pousse les répondants à exclure de l'information lorsqu'ils fournissent leur jugement comparativement à ce qui, dans d'autres, les incite à inclure l'information lorsqu'ils portent un jugement. Les progrès dans ce domaine nécessiteront une collaboration étroite entre les psychologues cognitifs et les méthodologistes d'enquête, ainsi que des travaux de laboratoire et sur le terrain.

Cependant, en dernière analyse, nous n'arriverons à la compréhension fondamentale du processus de question et de réponse que quand nous comprendrons comment la

Cette procédure permet d'estimer le comportement d'un groupe en réaction aux questions délicates, mais non celui d'un individu isolé. Donc, elle ne permet pas d'établir le lien entre les caractéristiques individuelles du répondant et le comportement individuel. Si l'échantillon est très grand, les caractéristiques du groupe peuvent être reliées aux estimations obtenues d'après les réponses randomisées. Ainsi, on pourrait examiner toutes les réponses des jeunes femmes et les comparer à toutes les réponses des hommes ou comparer les réponses des jeunes à celles des groupes d'âge plus âgés. Dans l'ensemble, une grande quantité d'information est malheureusement perdue lorsqu'on utilise la méthode à réponse randomisée.

Si, comparativement à d'autres méthodes, celle des réponses randomisées réduit considérablement la sous-déclaration de comportement indésirable, elle réduit à peine la surdéclaration des comportements souhaitables. Elle ne permet pas non plus d'éliminer entièrement la sous-déclaration des comportements indésirables (Bradburn et coll. 1979).

CONCLUSION

Le présent essai a pour but de présenter les grandes lignes d'une approche fondée sur la psychologie sociale de la compréhension du processus de question et réponse dans l'interview d'enquête. Cette approche s'inspire des théories de la sociologie, de la psychologie cognitive et de la linguistique pour présenter un cadre général d'étude des effets de réponse. Malheureusement, beaucoup d'éléments demeurent incertains ou inconnus.

La théorie du rôle social représente un bon point de départ pour la conceptualisation des relations sociales entre les chercheurs, les intervieweurs et les répondants, mais nous ignorons encore beaucoup de choses sur la façon dont ces rôles sont interprétés par les acteurs respectifs et sur la façon dont ils peuvent évoluer. Les préoccupations actuelles quant au respect de la vie privée, la confidentialité des données et la protection des sujets humains qui participent à la recherche sont en train de modifier dans des proportions inconnues la façon dont les répondants perçoivent les enquêtes et la recherche sociale. La technologie modifie la capacité des répondants à protéger leur vie privée et la capacité des chercheurs à protéger la confidentialité des données. Les taux de réponse sont à la baisse et des efforts énormes sont nécessaires pour convaincre les personnes échantillonnées de participer aux enquêtes. L'interview est de plus en plus fréquemment assistée par ordinateur, ce qui pourrait modifier la façon dont les répondants et les intervieweurs interagissent et la façon dont les répondants voient la situation d'interview.

Les processus cognitifs qui interviennent dans la formulation d'une réponse sont complexes et encore mal compris.

sociale, mais aucune ne permet de l'éliminer totalement et fiablement. La stratégie générale consiste à augmenter la distance sociale entre le répondant et l'intervieweur. On peut, pour cela, modifier le mode d'administration en éliminant ou en réduisant la présence de l'intervieweur. L'interview sur place assistée par ordinateur (IPAO), qui permet au répondant d'entrer directement les réponses aux questions délicates dans l'ordinateur dans le cadre d'une interview sur place permet aux chercheurs de combiner les avantages d'une interview sur place à ceux d'une question-naire à remplir soi-même. L'utilisation de l'IPAO avec fonction audio (IPAO-audio) qui permet au répondant d'écouter une voix lisant les questions, quoiqu'un peu plus coûteuse, est un moyen de contourner les problèmes de compétence en lecture et de langage qui pourraient se poser lorsque le répondant doit lire les questions sur un écran d'ordinateur.

La recherche sur les effets de mode indique générale-ment que le répondant soi-même à un questionnaire, parti-culièrement dans une situation anonyme de groupe, réduit au minimum, sans toutefois éliminer entièrement, le biais dû à la désirabilité. Les interviews réalisées par téléphone produisent généralement des résultats intermédiaires par rapport à l'interview sur place et l'auto-administration entièrement anonyme, quoique les résultats ne soient pas complètement convergents.

En plus d'augmenter la distance sociale entre l'intervieweur et le répondant en modifiant le mode d'adminis-tration, il existe des techniques permettant d'augmenter l'anonymité réelle ou perçue du répondant qui réduisent aussi le biais dû à la désirabilité sociale. Par exemple, le répondant peut placer ses réponses dans une enveloppe scellée et envoyer celle-ci par la poste à un bureau central, si bien qu'il sait que l'intervieweur ne peut pas voir ses réponses.

Une autre technique est celle dite de la réponse aléatoire, bien qu'il s'agisse plus correctement d'une technique à question aléatoire (Greenberg, Abul-Ela, Simmons et Horvitz 1969; Horvitz, Shah et Simmons 1967; Warner 1965). L'intervieweur pose deux questions, une délicate et l'autre non délicate. Les deux questions ont les mêmes réponses possibles, « oui » et « non ». La question qui est posée au répondant est déterminée par un mécanisme probabiliste, comme jouer à pile ou face ou utiliser une boîte en plastique contenant des billes de deux couleurs, par exemple rouges et bleues, en proportions différentes, par exemple 70 % de billes rouges et 30 % de billes bleues. La boîte est conçue de telle façon que, lorsque le répondant la secoue une bille rouge ou bleue vue uniquement de lui apparait dans la fenêtre de la boîte. Si la bille est rouge, il répond à la question délicate et si elle est bleue, il répond à la question non délicate. L'intervieweur ne sait pas à quelle question le répondant a répondu.

Formatage et vérification des réponses

répondants qui avaient estimé leur consommation de bière pour commencer (un item dont la fréquence est élevée), ont jugé le vin, le lait et le café comme étant des boissons moins typiquement allemandes que ceux qui avaient estimé la consommation de vodka pour commencer (un item à fréquence faible), ce qui témoigne d'un effet de contraste s'étendant aux trois boissons cibles. Cet effet de contraste s'est, par contre, pas manifesté lorsque la question précédente portait sur le contenu calorique de la bière ou de la vodka, parce que l'information activée par cette question était sans rapport avec le jugement au sujet de la typicité.

Après que les répondants ont formulé leurs réponses, il leur reste à les ajuster au format de réponse demandé par l'intervieweur. Il est rare, lors d'une enquête, qu'un chercheur permette aux répondants de répondre aux questions dans un format libre. Les questions ouvertes posent une multiplicité de problèmes, notamment le coût et la difficulté de la transformation des réponses libres en un format pouvant être traité quantitativement. Aujourd'hui, presque tous les questionnaires contiennent des questions fermées ou précodées.

La recherche théorique sur les options de réponse est moins avancée que l'étude des effets de l'énoncé de la question et du contexte. En général, les effets observés empiriquement sont considérés comme émanant de deux sources, à savoir la limite de la mémoire et l'élaboration cognitive stimulée par les options de réponse.

Les limites de la mémoire créent certains effets d'ordre parmi les options de réponses. La primauté et la récence sont deux effets bien décrits dans la littérature sur la mémoire. Lorsqu'il existe une série de stimuli visuels, ceux qui apparaissent les premiers dans la série sont mieux mémorisés que ceux qui surviennent plus tard (primauté). Lorsqu'il existe une série de stimuli en mode auditif, ceux qui surviennent en dernier dans la série sont mieux mémorisés (récence). Donc, il existe une interaction entre l'ordre dans lequel les stimuli sont présentés et le mode dans lequel ils sont présentés.

La littérature sur la recherche indique qu'il existe des effets persistants, quoique assez faibles dans les échantillons généraux, de primauté et de récence dus à la position sérielle des options de réponse et dépendant du mode de présentation. Les effets de primauté se manifestent quand les options de réponse sont présentées visuellement, par exemple au moyen de cartes lors d'une interview sur place, et les effets de récence se manifestent lors de l'interview téléphonique, lorsque les répondants doivent dépendre entièrement de leur mémoire auditive pour les options de réponse. Néanmoins, les études les plus récentes (Knauper 1999; Schwarz et Knauper 2000) révèlent que l'effet est en grande partie fonction de la capacité de mémoire et est fortement augmenté chez les répondants âgés dont la

limites de la mémoire. Lorsqu'une option de réponse est choisie dans l'esprit du répondant, ce dernier peut encore vérifier la réponse. Tel que mentionné plus haut, l'interview est une situation sociale et le répondant peut s'inquiéter de l'image qu'il projette. Il existe une foule de preuves que la désirabilité sociale est un aspect important du processus de réponse et que les réponses à des questions délicates peuvent être gravement faussées parce que le répondant n'est pas disposé à admettre qu'il a des comportements ou des attitudes pouvant le présenter à l'intervieweur sous un jour défavorable ou parce qu'il souhaite déclarer un comportement socialement souhaitable (Bradburn et Bradburn 1974). Plusieurs techniques existent pour réduire le biais dû à la désirabilité

lorsqu'une option de réponse est choisie dans l'esprit du répondant, ce dernier peut encore vérifier la réponse. Tel que mentionné plus haut, l'interview est une situation sociale et le répondant peut s'inquiéter de l'image qu'il projette. Il existe une foule de preuves que la désirabilité sociale est un aspect important du processus de réponse et que les réponses à des questions délicates peuvent être gravement faussées parce que le répondant n'est pas disposé à admettre qu'il a des comportements ou des attitudes pouvant le présenter à l'intervieweur sous un jour défavorable ou parce qu'il souhaite déclarer un comportement socialement souhaitable (Bradburn et Bradburn 1974). Plusieurs techniques existent pour réduire le biais dû à la désirabilité

mémoire est moins bonne et qui dépendent davantage de la présentation. Chez les répondants âgés, les effets de primauté/récence peuvent être assez importants, de l'ordre de 20 points de pourcentage (Schwarz et Knauper 2000). Chez les répondants plus jeunes, ces effets sont faibles.

Une théorie fascinante pour expliquer certains effets d'ordre de réponse observés lors de la réponse à une question est celle de l'élaboration cognitive. Cette théorie s'inspire des premiers travaux de Krosnick et Alwin (1987) et de la recherche cognitive sur la persuasion (Eagly et Chaiken 1993; Petty et Cacioppo 1986). Elle s'appuie sur l'hypothèse selon laquelle l'ordre et le mode de présentation des options de réponse ont une incidence sur la possibilité d'élaboration, à son tour, active des réflexions en réponse à la question et fournit les signaux d'extraction en réponse aux questions sur le comportement. Les options de réponse fournissent des signaux supplémentaires qui activent une gamme de pensées qui deviennent temporairement accessibles et peuvent faire partie intégrante du processus de formulation de la réponse. En effet, les options de réponse sont une partie essentielle de la question, mais elles peuvent être traitées plus tard que la question proprement dite.

L'hypothèse de l'élaboration cognitive suggère plusieurs prédictions complexes, dont un petit nombre seulement ont déjà été vérifiées. Un exemple pour lequel il existe un grand nombre de preuves est celui de l'interaction entre la position sérielle et le mode d'administration dans le cas de longues listes. L'effet de primauté évident dans le matériel présenté visuellement donne aux répondants le temps et le stimulus visuel pour réfléchir davantage aux options qui se situent au début de la liste avant de donner leur réponse. L'inhibition des premières options par la lecture des dernières et les effets de récence évidents dans les listes présentées selon un mode auditif donne à penser que les dernières options peuvent être traitées plus profondément du point de vue cognitif. Ces effets sont plus robustes que les effets de primauté et de récence qui semblent dépendre davantage des simples

parce que le jugement requis pour répondre à la question est fondé sur l'information intégrée dans la représentation utilisée. Si l'information est positive, le jugement sera plus positif, tandis que si elle est négative, il sera plus négatif. La grandeur de l'effet dépend de la quantité et de la force de l'information temporairement accessible.

Les questions antérieures peuvent activer des pensées qui sont alors incluses dans la représentation des sujets sur lesquels portent les questions ultérieures. L'effet d'une question donnée diminue à mesure que le nombre d'autres questions contextuelles augmente. Ainsi, répondre à une question au sujet du bonheur conjugal a un effet prononcé sur les réponses aux questions subséquentes au sujet de la satisfaction générale à l'égard de la vie lorsque le mariage du répondant est le seul domaine spécifique de la vie sur lequel des questions sont posées. Par contre, si l'on pose des questions sur les loisirs et sur les emplois en plus de questions sur le mariage avant de poser les questions sur la satisfaction à l'égard de la vie, l'effet est réduit significativement (Schwarz, Strack et Mai 1991).

L'information qui est exclue plutôt qu'incluse dans la représentation temporaire de la cible donne lieu à un effet de contraste. Dans ce cas, si l'information exclue est positive, le jugement deviendra plus négatif, tandis que si elle est négative, le jugement deviendra plus positif. De la même façon, la taille de l'effet dépend de la quantité et de la force de l'information accessible temporairement. En effet, de l'information exclue est soustraite de la représentation de l'objet de l'attitude.

Cependant, l'information peut jouer un rôle supplémentaire dans la formulation des jugements. En plus d'être exclue de la représentation de la cible, l'information peut être utilisée pour construire une norme ou un encrage d'échelle. Dans ce cas, nous parlons d'effet de contraste fondé sur une comparaison. Ici, l'effet est causé non pas tant par la soustraction de l'information exclue de l'évaluation de la cible de l'attitude que par la comparaison de la cible à une norme donnée ou à son évaluation sur une échelle.

Celui de ces processus qui produit un effet de contraste détermine si cet effet est limité au seul objet ou généralisé à tous les objets connexes. Si l'effet de contraste est fondé sur une simple soustraction, il est limité à cette cible particulière. Par contre, s'il est fondé sur une comparaison, il peut apparaître dans chaque jugement où la norme de comparaison est pertinente.

Une étude réalisée par Schwarz, Muenkel et Hippler (1990) fournit un exemple d'effet de contraste fondé sur l'utilisation de l'information provenant de questions antérieures. On a demandé aux répondants d'évaluer, pour un certain nombre de boissons, la mesure dans laquelle elles étaient « typiquement allemandes ». Quand cette question était précédée par une question sur la fréquence à laquelle les Allemands boivent de la bière ou de la vodka, des effets de contraste sont apparus dans l'évaluation de la typicité. Les

durant la période visée augmentera à mesure que les événements remontent plus loin dans le temps. Cette surdéclaration des événements survenus en dehors de la période visée par la question ne sera pas compensée par une sous-déclaration des événements survenus dans le court terme, parce que les événements qui n'ont pas encore eu lieu ne peuvent pas être déclarés. Puisqu'il n'existe aucun événement compensateur remémoré comme ayant eu lieu en dehors de la période délimitée, mais à l'autre extrémité, c'est-à-dire l'avant, le résultat est une surdéclaration nette. (Pour une explication complète du modèle, consulter Huttenlocher, Hedges et Bradburn 1990).

Formulation d'une réponse

En tenant compte de l'information qui est activée par les signaux fournis par les questions et le contexte dans lequel elles sont posées et qui est extraite de la mémoire, les répondants doivent formuler une réponse à la question. Certains renseignements sont facilement accessibles. Ainsi, si les questions portent sur des sujets bien mémorisés, comme les dates de naissance ou l'état matrimonial, ou sur des sujets pour lesquels existe une position déjà bien précisée, le répondant peut extraire les réponses directes-ment. Elles surgissent, telles quelles, entièrement formées, de la mémoire et peuvent être déclarées directement. Aussi qu'alloons-nous cette forme d'information de chroniquement accessible.

Par ailleurs, si les questions portent sur un comportement auquel le répondant n'a pas pensé récemment et dont il ne se souvient pas bien, ou sur des attitudes qui n'ont pas fait l'objet de mûre réflexion ou de discussion, le répondant doit construire des réponses sur le tas, en utilisant toute l'information provenant de toute source disponible dans la mémoire de travail. Ce processus de construction utilise non seulement l'information chroniquement disponible, mais aussi, fait important, l'information qui est temporairement accessible parce qu'elle a été activée par la question proprement dite, les signaux contextuels, les questions antérieures ou tout autre aspect de la situation d'interview.

Il existe plusieurs processus cognitifs généraux qui sont des stratégies puissantes pour traiter l'information efficacement. L'assimilation et le contraste sont deux de ces processus fondamentaux qui influent sur les comportements. Dans l'étude de la perception, l'assimilation s'entend de la tendance à percevoir les stimuli comme étant plus semblables qu'ils ne le sont réellement. Le contraste, quant à lui, s'entend de la tendance à percevoir les stimuli comme étant plus différents qu'ils ne le sont réellement. L'application de ces principes à la réponse à un questionnaire d'enquête amène à ce que l'on a appelé le modèle d'inclusion/exclusion (Schwarz et Bless 1992; Sudman et coll. 1996). L'information qui est intégrée dans la représentation temporaire que le répondant forme de la cible de la question donnera lieu à des effets d'assimilation,

filie (Brewer 1986, 1994). Pour que des activités soient gardées en mémoire, elles doivent être saisies. Autrement dit, elles doivent être comprises dans un système de signification, habituellement linguistique, qui fait intervenir la connaissance des activités passées et la connaissance générique au sujet d'événements de type similaire, ainsi que les connaissances spécifiques au sujet de l'événement proprement dit et le contexte dans lequel il survient. Cet ensemble complexe d'information utilisée pour comprendre l'événement devient partie intégrante de la mémoire de l'événement. Le processus de compréhension détermine comment les souvenirs sont encodés.

L'information, comme l'énoncé de la question et tout matériel explicatif à la disposition des répondants au moment où on leur demande de se rappeler d'un événement, joue le rôle de signal d'extraction. Tout mot, image, émotion, etc., qui active ou oriente le processus de recherche dans la mémoire est un signal d'extraction. Si ces signaux ne spécifient pas le type d'événement, par exemple, les visites chez le pédicure, alors les types d'événement doivent être insérés avant que la recherche débute. Cette inférence peut provenir de l'énoncé de la question ou du contexte plus général dans lequel la question est posée, y compris les questions précédentes et le matériel d'introduction à l'enquête.

L'extraction est un processus actif qui est facilité par des signaux compris dans la question qui active les voies d'association menant à l'information souhaitée. Comme, aussi bien dans la mémoire épisodique que dans la mémoire sémantique, l'information est encodée de nombreuses façons différentes, les signaux contenus dans la question ou dans le contexte dans lequel celle-ci est posée, y compris les questions antérieures, peuvent faciliter ou restreindre l'activation et produire une meilleure ou une moins bonne extraction.

L'extraction prend du temps. Un résultat empirique clair est que, si l'on donne plus de temps aux répondants pour répondre aux questions, les déclarations sont plus exactes, particulièrement pour les questions relatives au comportement. Mais le temps n'est pas tout. Les souvenirs d'événements survenus durant la vie d'une personne semblent être organisés en séquence d'événements (Barsalou 1988), par exemple, des vacances d'été et une hospitalisation, qui sont aux répondants des signaux qui leur rappellent la séquence que d'essayer de leur faire extraire l'information au sujet d'un événement spécifique. Ainsi, dans les questions sur la consommation d'alcool, donner des exemples du genre de situation dans lesquelles on pourrait boire augmente les déclarations de consommation.

Les exemples sont une aide importante à la mémoire-ration, mais il ne s'agit pas d'une panacée. Donner aux répondants une liste de magazines qu'ils pourraient avoir lus améliore les déclarations sur la lecture; une liste de types

Quand nous réfléchissons à l'extraction, nous pensons principalement à l'oubli de l'information pertinente, ou l'impossibilité de l'extraire. Néanmoins, il peut arriver que des renseignements inconnus soient extraits, ce qui se traduit par un comportement de surdéclaration. L'exemple le mieux connu est le phénomène observé par Neter et Waksberg (1964), baptisé « télescopage ». C'est-à-dire la remémoration d'événements qui ont eu lieu à une autre période que celle sur laquelle la question. Le télescopage se produit lors de la réponse à des questions concernant le comportement durant des périodes précises, comme : « Combien de fois avez-vous consulté le médecin au cours des six derniers mois? ». En analysant les données de la Consumer Expenditure Survey, Neter et Waksberg ont constaté que, si les répondants font une déclaration sur des achats lors de diverses périodes de référence, il se produit une surdéclaration systématique des achats due au fait de déclarer ceux survenus durant une période antérieure, comme s'ils avaient été faits durant la période visée par la question. Bien que ce phénomène ait été observé lors de plusieurs études, aucune explication cognitive n'a été avancée jusqu'à récemment.

Le souvenir du moment des événements devient d'autant plus incertain que ces événements remontent loin dans le temps, même si il n'existe aucun biais systématique dans les déclarations. Le télescopage résulte de la conjugaison de deux processus, à savoir l'arrondissement et l'établissement des bornes. L'arrondissement s'entend du fait que les répondants arrondissent leurs estimations du moment où les choses se sont passées en périodes de plus en plus longues à mesure que l'événement remonte plus loin dans le temps. Par exemple, une personne se souviendra que les événements ont eu lieu « il y a dix jours », puis elle l'arrondira à des périodes telles qu'il y a dix jours, deux semaines, quatre semaines, trois mois ou six mois. L'établissement des bornes s'entend de l'aspect de la question qui limite la période visée par la déclaration, par exemple les six derniers mois. L'établissement de ces bornes a pour effet de tronquer la déclaration des événements qui, selon les souvenirs de la personne ont eu lieu il y a plus de six mois. Puisque la variance dans le souvenir des dates des événements augmente à mesure que l'événement remonte plus loin dans le temps, le nombre d'événements dont le répondant se souviendra avoir eu lieu

même pour des questions utilisant ce genre de termes

La compréhension débute par un processus perceptuel

d'interprétation d'une chaîne de sons ou de symboles écrits,

comme des mots, en un langage que comprennent les

répondants. La chaîne de mots est « analysée » en unités

syntactiques compréhensibles, c'est-à-dire que la signifi-

cation encodée dans les unités linguistiques est extraite par

un processus qui reste mal compris. Nombre de problèmes

de compréhension sont causés par des ambiguïtés dues au

fait que certains mots ont plusieurs significations (ambiguïté

lexicale) ou peuvent être utilisés de plusieurs façons

(ambiguïté structurelle). Par exemple, la question « Où est la

table? » est lexicalement ambiguë, parce que le mot

« table » peut désigner un objet sur lequel on peut placer des

choses ou un ensemble de nombres disposés sur une feuille

de papier. La phrase « J'ai lu la critique de Chomsky » est

structuellement ambiguë. L'interprétation varie selon que

« Chomsky » est compris comme étant l'auteur ou l'objet de

la critique. Les ambiguïtés structurelles peuvent être

résolues en énonçant soigneusement les questions. Les

ambiguïtés lexicales, par contre, sont inhérentes au langage

et habituellement résolues d'après le contexte dans lequel la

phrase apparaît.

Le contexte joue un rôle important non seulement pour

ce qui est de résoudre les ambiguïtés, mais aussi pour

faciliter l'interprétation de la signification de mots inconnus.

Par exemple, Schuman et Presser (1981) ont constaté

qu'une question au sujet de la Monetary Control Bill, un

projet de loi obscur, était interprétée comme faisant allusion

à une mesure anti-inflationniste lorsqu'elle était posée après

une série de questions au sujet de l'inflation, mais qu'elle

était interprétée comme faisant allusion à des contrôles du

transfert international d'argent quand elle était posée après

des questions ayant trait à la balance des paiements.

Le mécanisme psychologique qui sous-tend ces types

d'effets contextuels est l'amorçage. Afin d'interpréter la

chaîne de sons ou de symboles écrits, nous devons faire

appel à notre mémoire sémantique qui contient la réserve

d'informations linguistiques qui nous permet de comprendre

les langages que nous connaissons. Puisqu'il s'agit d'un

grand entrepôt de connaissances, l'extraction de l'informa-

tion prend du temps et certains éléments seront plus

facilement accessibles que d'autres. Les fragments d'informa-

tion qui ont été activés récemment sont plus facilement

accessibles et seront utilisés pour commencer pour inter-

prêter ce qui est dit ou lu. L'amorçage active les pensées ou

« schémas », c'est-à-dire les pensées organisées au sujet

d'objets ou de concepts, de sorte qu'elles soient plus

accessibles à la conscience et, donc, intervenient plus

facilement dans l'interprétation des questions. Dans

l'exemple susmentionné, les questions antérieures ont

amorçé soit les pensées au sujet de l'inflation soit celles au

sujet des flux internationaux d'argent, si bien que, lorsqu'on

pose une question au sujet du concept inconnu de Monetary

Control Bill, les pensées qui ont été amorcées viennent plus

rapidement à l'esprit et influencent l'interprétation des mots.

Différentes significations peuvent être accessibles

différemment pour divers répondants en raison de la

fréquence différente à laquelle ils les emploient dans leur vie

quotidienne. Par exemple, Billiet (cité dans Bradburn 1992,

page 317) observe qu'à la question « Combien d'enfants

avez-vous? », certains répondants ont indiqué des nombres

compris entre 20 et 30. Un examen plus approfondi des

données a montré que ces répondants étaient des

enseignants qui ont interprété la question comme faisant

allusion aux enfants dans leur classe, la signification qui

était la plus accessible dans leur mémoire.

Extraction de l'information

Quand une question a été comprise, les répondants doivent

extraire de leur mémoire l'information nécessaire pour

répondre à la question. Dans presque tous les cas, cela

signifie extraire l'information de la mémoire à long terme.

Si la question a trait à un comportement, l'information

pertinente est vraisemblablement entreposée dans la

mémoire épisodique. Si la question a trait à des attitudes,

l'information pertinente est vraisemblablement entreposée

dans la mémoire sémantique, mais peut nécessiter

l'extraction de certaines informations à partir de la mémoire

épisodique.

La remémoration est un processus en vertu duquel une

recherche est effectuée dans l'entrepôt de la mémoire afin

de retrouver l'élément particulier recherché. Si nous nous

représentons la mémoire comme un énorme entrepôt, il est

clair qu'elle doit présenter un certain degré d'organisation

afin que nous puissions en extraire des éléments. Au même

titre que nous devons étiqueter les fichiers avant de les

placer dans les tiroirs d'un classeur, nous devons apposer

une forme ou l'autre d'étiquette à l'information stockée

dans l'entrepôt de la mémoire. Le processus d'étiquetage,

souvent appelé « encodage », fait référence à divers aspects

de l'information ou de l'expérience, y compris le ton

émotionnel associé à l'élément au moment où nous l'avons

entreposé dans la mémoire afin de pouvoir l'extraire. (Pour

une discussion plus complète des modèles de la mémoire,

voir Tourangeau et coll. 2000, chapitre 3).

Barsalou (1988) a proposé une théorie qui fournit un bon

cadre pour comprendre la façon dont l'information au sujet

d'événements personnels est gardée en mémoire. Il fait

remarquer que l'information au sujet d'activités ou d'évène-

ments types figurent dans la mémoire épisodique inclut non

seulement les événements spécifiques, mais aussi des

connaissances générales approfondies au sujet des

événements, c'est-à-dire ayant une image mentale générale

de certains types d'activité, comme les visites chez un

pédiateur, plutôt que l'image d'un événement particulier,

comme consulter le Dr. Larose au sujet du rhume de votre

Processus cognitifs dans la réponse aux enquêtes

Répondre aux questions d'une enquête demande un travail cognitif considérable au répondant. Les fondements des progrès récents dans la compréhension des processus de réponse aux enquêtes tiennent en grande partie à l'application de modèles de traitement de l'information au processus de question et réponse. Il nous reste, certes, encore beaucoup de recherche à faire avant que nous puissions comprendre entièrement et de façon détaillée comment le cerveau traite l'information, mais l'approche générale est suffisamment bien acceptée pour servir de base à une meilleure compréhension du processus de réponse.

Conceptuellement, l'esprit peut être considéré comme un grand système de traitement de l'information comprenant une série de systèmes composants. Les sensations physiques auditives et visuelles entrent dans le système au niveau du registre sensoriel. Ce dernier a des limites de capacité si bien que seule une partie de l'information est transférée dans la mémoire à court terme. L'attention joue un grand rôle dans la détermination de ce qui est entreposé dans cette mémoire. L'attention est une fonction d'un système superviseur qui active et contrôle le système de traitement de l'information de façon fort semblable aux programmes qui rendent les ordinateurs opérationnels. Le système superviseur contrôle l'entière du système grâce à des objectifs et des plans qui sont organisés selon les priorités d'action.

L'entrepôt du système est la mémoire à long terme dont la capacité est très grande. La mémoire de travail est le système dans lequel a lieu la réflexion. Ici, l'activité s'appuie sur la mémoire à court terme et des extractions à partir de la mémoire à long terme. La mémoire à court terme a une capacité limitée, mais son accès est rapide, tandis que la mémoire à long terme a une grande capacité, mais son accès est relativement lent. La mémoire à long terme semble posséder deux sous-systèmes assez distincts, la mémoire sémantique et la mémoire épisodique, bien que cette distinction ne soit pas universellement reconnue. La mémoire sémantique s'entend de la mémoire associée au vocabulaire, à la structure du langage, aux règles et aux connaissances abstraites, tandis que la mémoire épisodique s'entend de la mémoire d'événements qui ont lieu dans le temps et dans l'espace.

L'information est représentée comme une liste de caractéristiques ou de concepts qui sont reliés en réseaux. Elle est gardée en mémoire dans des structures qui sont organisées hiérarchiquement, les concepts les plus généraux occupant un niveau plus élevé dans la structure que les occurrences plus discrètes du concept ou les caractéristiques distinctes. On utilise parfois le terme « schéma » pour désigner les structures communes et (ou) exposées à un sur-apprentissage plus grandes et plus complexes qui organisent nos pensées sur des sujets familiers et peuvent être extraites comme un tout plutôt que sous forme de fragments.

Compréhension

Afin de répondre à une question, le répondant doit d'abord comprendre ce qu'on lui demande. L'objectif du chercheur est que les répondants comprennent la question de la même façon que lui. Cet objectif est très difficile à atteindre, à cause des nombreuses subtilités et ambiguïtés du langage. En effet, Belson (1981), qui a étudié en profondeur la compréhension de termes communs, comme « jour de semaine », « enfant », « régulièrement » et « proportion » a constaté que la mauvaise compréhension est généralisée, se produisant selon une séquence ordonnée.

Le langage est le moyen principal de communication de l'information et, pour être disponible pour la communication, l'information doit donc être associée à un code linguistique. La relation exacte entre le langage et la pensée, ainsi que la question de savoir si toutes les pensées ont ou non une représentation verbale continuent de faire l'objet de débat. Cependant, il est clair que la signification est encodée d'une certaine façon dans le langage et que ce code joue un rôle important dans l'acquisition, le stockage et l'extraction de l'information. L'émotion pourrait aussi faire partie du code, mais son rôle est encore mal compris.

Les structures de connaissance facilitent et contraignent les mécanismes d'activation de l'esprit. Ce qui vient à l'esprit, c'est-à-dire, devient conscient, est limité et est le résultat de l'activation des réseaux. L'activation est rapide, mais suit des voies déterminées par la façon dont l'information est encodée. Le codage place l'information dans des catégories particulières et structure les voies par lesquelles l'information sera extraite. Les signaux sont des stimuli qui sont reliés aux codes et stimulent l'activation des réseaux. Le temps que prend une personne pour répondre à un stimulus (temps de réaction) est souvent utilisé en recherche comme un indice de la façon dont l'information est codée.

Il existe plusieurs modèles du processus de question et réponse (Cannell, Miller et Oksenberg 1981; Strack et Martin 1987; Tourangeau et Rasinski 1988; Sudman et coll. 1996) qui, s'ils diffèrent dans les détails, concordent généralement pour une série de processus par lesquels passent les répondants quand ils répondent aux questions. Ces processus sont 1) comprendre la signification de la question, 2) extraire l'information pertinente, 3) formuler une réponse, 4) formater et vérifier la réponse de sorte qu'elle satisfasse aux exigences de l'intervieweur et de la façon dont veut se présenter le répondant. Bien que ces processus soient conceptualisés comme une séquence linéaire, il est reconnu qu'en fait, ils ont lieu dans le courant de la conversation et que plusieurs d'entre eux peuvent se dérouler parallèlement ou en cycle aller-retour rapide. Toutefois, pour étudier les processus de question et réponse, il est plus simple de les considérer comme étant distincts et

d'être un bon répondant que de simplement en finir avec l'interview. Donc, elle pourrait prendre moins de temps pour réfléchir aux questions, faire moins d'effort pour se satisfaire d'une réponse honnête plutôt qu'un « je ne sais pas », voire même une réponse fausse. Certains intervieweurs m'ont dit qu'ils ont souvent le sentiment que les réponses données par les personnes qu'ils ont convaincues de participer à une interview après qu'elles aient essayé de refuser la conversation sont moins valides que celles fournies par les personnes qui participent de bon gré. Les efforts supplémentaires pour obtenir un taux élevé de réponse risquent, en fait, de produire de moins bonnes données.

Les répondants peuvent aussi mal interpréter la nature de l'interview d'enquête, vouloir simplement la convertir en une conversation sociale ou ne pas être très aptes à la conversation, autrement dit ne pas suivre les maximes griciennes et, donc, engager une conversation « inefficace ». Ces conversations sont caractérisées par des à-côtés ou des changements de sujet fréquents, des commentaires sur des sujets n'ayant que peu ou pas de rapport avec la question considérée, le compte rendu d'anecdotes personnelles évoquées par certains aspects de la question, ou la simple répétition de commentaires. Dans de tels cas, l'intervieweur doit poliment, mais fermement, enseigner au répondant les règles de la conversation et le guider afin qu'il se concentre sur les questions de l'interview. Les intervieweurs chevronnés arrivent à contrôler expertement la conversation et, grâce à un renforcement sélectif, à modeler le comportement du répondant de sorte qu'il suive les maximes griciennes.

Brièvement, les interviews ont lieu dans des contextes sociaux dont la structure est régie par des attentes et des normes partagées par la société. Ces normes peuvent différer d'une société à l'autre, voire même d'une sous-culture à l'autre dans une même société, mais elles ont de puissants effets sur la façon dont les interviews sont menées et dont les questions sont interprétées. La violation des attentes ou des normes peut donner lieu à des « effets » susceptibles d'être interprétés comme des erreurs du point de vue du chercheur. Si elles sont comprises, les normes et attentes peuvent être utilisées pour éviter les problèmes ou pour atténuer les effets.

Des données pourraient aussi être recueillies auprès des intervieweurs sur la mesure dans laquelle l'interview s'est écartée du modèle décrit plus haut. Peu de recherche a été faite en vue d'évaluer la qualité des interviews de ce point de vue, mais l'examen de la diminution de la validité des données à mesure que les conditions de l'interview s'écartent du modèle idéal pourrait constituer un domaine d'étude fructueux dans l'avenir.

L'un des effets d'ordre les mieux décrits dans les enquêtes survient lorsque des questions de divers niveaux de spécificité sont posées ensemble. Lorsqu'une question est générale, par exemple, « Tous les éléments pris ensemble, dans quelle mesure vous sentez-vous heureux(se) ces jours-ci ? » et que l'autre est spécifique, par exemple « Dans quelle mesure votre mariage est-il heureux ? », les réponses à la question générale sont influencées par l'ordre des questions, tandis que les réponses à la question plus spécifique ne le sont pas. L'effet semble être le résultat du rôle joué par la maxime de pertinence. Lorsque la question générale est posée pour commencer, elle est interprétée comme elle est destinée à l'être, c'est-à-dire que les répondants devraient inclure tous les aspects de leur vie pour évaluer leur degré de bonheur. Lorsque la question générale est posée après la question particulière au sujet du bonheur conjugal, la maxime de pertinence indique aux répondants qu'ils devraient exclure le mariage des éléments pris en considération, puisqu'ils ont déjà fait une déclaration à ce sujet. Donc, même si la question vise littéralement « tous les éléments pris ensemble », elle est interprétée comme signifiant « tous les éléments, sauf ceux au sujet desquels des questions ont déjà été posées ». Seuls les éléments au sujet desquels aucune question n'a encore été posée sont

que se passe-t-il si les normes susmentionnées ne sont pas acceptées dans l'interview, parce que le répondant les rejette ou qu'il redéfinit le rôle du répondant ou encore qu'il n'observe pas les maximes de la conversation? Naturellement, la forme la plus facile de rejet du répondant consiste à refuser tout bonnement de participer à l'interview. Toutefois, il arrive qu'une personne échantillonnée devient « un répondant réticent » parce qu'elle se sent forcée de participer à l'étude à cause des procédures de suivi, parce qu'elle n'aitime pas dire non à une demande insistante d'une autre personne ou à cause d'une autre raison. Le cas échéant, cette personne pourrait se soucier nettement moins

rencontre et 3) les coûts et les avantages pour le répondant, s'il accepte de participer à l'interview. L'interaction est donc perçue comme étant neutre, ayant un but et en valant la peine. Comme toute autre interaction sociale instructure, elle est régie par les normes relatives à ce genre d'interaction.

Quelles sont les normes qui importent pour l'interview?

La première est le respect mutuel des individus, particulièrement le respect de la vie privée du répondant. Ce principe a pris une grande importance dans le contexte de la protection des participants aux études, parce que, lors d'un certain nombre d'études biomédicales, la nature volontaire de la participation n'a pas été explicitée clairement. Pour les travaux de recherche à haut risque, le consentement écrit du participant est désormais requis. Toutefois, dans le cas de l'interview d'enquête, le contexte de la demande de participation est tel qu'il est facile au répondant de refuser de participer d'enquête, le contexte de la demande de participation est tel qu'il est facile au répondant de refuser de participer. Demander un consentement écrit pourrait, en fait, superflus. S'il ne le souhaite pas, et son consentement écrit est écrit ne fait normalement pas partie d'une conversation entre inconnus qui ont établi que l'interaction n'est pas menaçante.

Une deuxième norme importante est l'honnêteté. L'obligation d'être honnête est incluse dans le rôle des deux parties. Pour l'intervieweur, cela signifie communiquer au répondant des faits pertinents quant au but de l'interview, ce qu'il est attendu du répondant, par exemple combien de temps l'interview durera, s'il faudra consulter des documents, si les questions risquent d'être délicates, etc., et répondre à toutes questions posées par le répondant. Si la fourniture de certains renseignements au début de l'interview, comme l'identité de l'organisme qui patronne l'enquête, risque d'introduire un biais dans les réponses, l'information peut être donnée à la fin de l'interview.

Le but de l'interview est d'obtenir l'information requise par le chercheur. Le rôle de l'intervieweur est d'obtenir l'information souhaitée et le questionnaire est l'instrument principal pour accomplir cette tâche. Un questionnaire bien conçu facilite la tâche de l'intervieweur et réduit au minimum le nombre de questions auxquelles celui-ci doit répondre concernant la signification des questions. Bien que les intervieweurs doivent recevoir une formation concernant le but des questions et leur signification, ils peuvent devenir une source de variance non contrôlée s'ils doivent interpréter les questions pour un grand nombre de répondants. Les intervieweurs doivent être sensibles aux indices signalant que les répondants comprennent mal les questions et prendre des mesures pour corriger la situation. Si les intervieweurs doivent intervenir fréquemment, le questionnaire est mauvais.

Si les répondants acceptent le rôle et conviennent de participer à l'interview, ils ont l'obligation, en vertu de la norme d'honnêteté, de répondre aux questions aussi

exactement et complètement que possible. Toutefois, cette norme peut être en conflit avec le désir général des individus de faire bonne impression et de se présenter sous le jour le plus favorable. Dans de nombreuses enquêtes, nous posons des questions au sujet de comportements éventuellement embarrassants, délicats, voire même illégaux, ou des attitudes impopulaires. L'intervieweur et le questionnaire jouent tous deux un rôle important dans la minimisation de ce conflit et renforcent la forme d'honnêteté. Néanmoins, selon des données empiriques, même les intervieweurs les mieux formés et les meilleures techniques de conception de questionnaire permettent rarement d'empêcher une certaine surdéclaration des comportements et attitudes socialement souhaitables ou la sous-déclaration des attitudes et comportements indésirables (voir Bradburn, Sudman et collaborateurs 1979; Wendland et Smith 1993).

Les données d'enquête sont recueillies en vertu d'une norme rigoureuse de confidentialité. Elle est si rigoureuse que même si elle n'est pas rendue explicite, les répondants s'attendent à ce que l'information provenant d'interviews qui ont la forme d'enquêtes scientifiques, comme les sondages d'opinion ou les enquêtes sur les attitudes des recrues désignées sous le vocable d'enquête) ou du « frugging » (campagnes de collecte de fonds désignées sous le vocable d'enquête) menacent de miner la confiance du public dans les enquêtes et de contribuer à la hausse du taux de refus de participer. À moins que les données soient recueillies aux termes de « lois protectrices » ou de certificats de confidentialité qui ont force de loi, les promesses de confidentialité peuvent être compromises par des activités d'application de la loi.

Les linguistes ont également remarqué que les conversations sont fondées sur un principe de « coopération » qui est concrétisé par quatre maxims. La maxime de qualité prescrit aux locuteurs d'être honnêtes et de ne pas dire des choses dont ils n'ont pas la preuve. La maxime de relation indique que les énonciations sont en rapport avec le sujet de la conversation en cours. La maxime de quantité exige que les locuteurs ne se répètent pas et rendent leur participation à la conversation aussi informative que possible. La maxime de la façon la plus intelligible possible, s'expriment de la façon la plus intelligible possible. Donc, selon Grice, il est attendu des locuteurs qu'ils soient honnêtes, pertinents, informatifs et intelligibles.

Ces maxims s'appliquent également aux conversations informelles et aux interviews ayant la forme d'une catégorie spéciale de conversations. Donc, les questions posées par

phénomènes objectifs ainsi que subjectifs. Les résultats du colloque ont été publiés dans Jabine et coll. (1984).

Le dernier cas de travaux indépendants pouvant être considéré à l'origine de ce domaine d'étude a été la conférence organisée par Norbert Schwarz et ses associés en Allemagne. L'article le plus influent émanant de cette conférence a sans doute été le modèle proposé par Strack et Martin (1987) « Thinking, judging and communicating: A process account of context effects in attitude surveys ». Les résultats de la conférence sont publiés dans Hippler, Schwarz et Sudman, Social Information processing and survey methodology (1987).

Durant les années qui ont suivi, un flot de travaux ont permis d'élaborer davantage le programme de recherche issu des premiers colloques. Une partie des travaux publiés par le Social Science Research Council sont consacrés à des questions relatives à la compréhension des erreurs de réponse dans les déclarations relatives aux comportements (Biderman 1980). Cet atelier a notamment eu pour effet d'inciter certains psychologues cognitifs à se lancer dans l'étude en laboratoire des problèmes que posent les questions d'enquête. L'un des tout premiers articles publiés, intitulé « Since the eruption of Mt. St. Helens has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events » (Loftus et Marburger 1985), démontre expérimentalement la valeur de l'utilisation d'événements-repères pour améliorer la qualité du datage des événements dans les déclarations faites lors des enquêtes.

Le deuxième événement marquant a été l'établissement par le Committee on National Statistics d'un groupe d'experts sur la mesure des phénomènes subjectifs. Ce groupe d'experts a produit un ouvrage en deux gros volumes contenant une revue de nombreux travaux de recherche sur les effets de réponse qui interviennent dans la mesure des phénomènes subjectifs. Cet ouvrage est venu compléter les travaux réalisés lors des séminaires antérieurs sur la mesure du comportement ou de phénomènes plus « objectifs » (Turner et Martin 1982).

Un élan important a été donné en 1983, quand le Committee on National Statistics a organisé, grâce à l'appui financier du NSF, un colloque de six jours à St. Michaels, Maryland, sur les aspects cognitifs de la méthodologie d'enquête. Il a donné lieu à la présentation de deux articles, intitulés « Potential contributions of cognitive research to survey questionnaire design » (Bradbourn et Danis 1984) et « Cognitive science and survey methods » (Tourangeau 1984) examinant comment les progrès récents en psychologie cognitive pourraient s'appliquer à la méthodologie d'enquête et comment les progrès en méthodologie d'enquête pourraient contribuer à l'évolution future de la psychologie cognitive. Extraordinairement fructueux, le colloque a donné naissance à un tout nouveau domaine de recherche en méthodologie d'enquête telle qu'appliquée aux

Une approche conceptuelle de l'interview d'enquête

Une interview d'enquête est une interaction sociale structurée entre deux personnes qui jouent des rôles distincts – l'intervieweur et le répondant. On l'a décrite comme étant une « conversation ayant un but » (Bingham et Moore 1934). Succinctement, le but consiste à obtenir les réponses à une série de questions. Dans les enquêtes scientifiques, les questions sont habituellement intégrées dans un questionnaire structuré conçu par une tierce partie, à savoir le chercheur. Ce type d'activité d'enquête est celui que je considérerai, quoique l'analyse puisse être étendue à d'autres formes d'interviews moins structurées.

Comme toutes les interactions sociales entre personnes appartenant à la même culture, des règles implicites influencent la façon dont les participants se comportent. Certaines de ces règles sont générales et s'appliquent à toutes les interactions sociales entre pairs sociaux; certaines sont générales pour le type particulier d'interaction que nous appelons l'interview d'enquête; certaines sont générales pour l'enquête; et certaines sont particulières et ne s'appliquent qu'à l'interview en question. Donc, nous concevons ces règles comme étant organisées hiérarchiquement, des plus générales, s'appliquant à toutes les interviews d'enquête, aux règles propres à une interview particulière.

Au niveau le plus général, l'interaction est régie par les règles s'appliquant aux interactions volontaires entre inconnus. Une interaction est initiée par l'une des parties, l'intervieweur, qui doit préciser la nature de la rencontre. Les éléments importants qui doivent être établis sont 1) que l'interaction n'est pas menaçante, autrement dit que l'intervieweur ne nuira pas au répondant, 2) le but de la

Comprendre le processus de question et réponse

NORMAN M. BRADBURN¹

RÉSUMÉ

Les statisticiens d'enquête savent depuis longtemps que le processus de question et réponse est une source d'effets de réponse qui contribuent à une erreur de mesure non aléatoire. Au cours des deux dernières décennies, l'application des concepts de la psychologie sociale et cognitive à l'étude du processus de question et réponse a permis de faire de grands pas vers la compréhension de ces sources d'erreur. Le présent essai a pour but de passer en revue le développement de ces approches, de discuter de l'état actuel de nos connaissances et de proposer certaines priorités de recherche pour l'avenir.

MOTS CLÉS : Erreurs de mesure; effets de réponse; psychologie cognitive; conception du questionnaire.

mesure dans laquelle les chercheurs comprennent les

réponses.

Même si cette estimation est trop pessimiste, nous avons

à résoudre un problème difficile d'erreur de mesure qui

plutôt que du plan de sondage ou de l'exécution de

l'enquête. L'existence de cette source d'erreur de mesure est

reconnue depuis la réalisation des premières enquêtes

scientifiques, autrement dit, depuis l'élaboration de la

théorie de l'échantillonnage et de son application aux popu-

lations humaines. Contrairement à la théorie de l'échantil-

lonnage, qui repose sur des principes mathématiques bien

établis, la compréhension de l'erreur de mesure due au

processus de question et de réponse n'a pas, jusqu'à

récemment, été fondée sur la compréhension théorique de la

communication et de la cognition humaines. Cette situation

est sur le point de changer.

Au cours des deux dernières décennies, des progrès

importants ont été réalisés dans la conceptualisation de

l'interview d'enquête, grâce à l'application de concepts

empruntés à la psychologie sociale et cognitive (Fabine,

Straf, Tanur et Tourangeau 1984; Sudman et Bradburn

1974; Sudman, Bradburn et Schwarz 1996; Tourangeau,

Rips et Rasinski 2000). Dans le présent essai, j'examinerai

brèvement l'élaboration de ces approches, je discuterai de

l'état actuel de notre connaissance du processus de question

et réponse et je proposerai certaines priorités de recherche

pour l'avenir.

Un peu d'histoire

La collaboration entre les psychologues cognitifs et les spécialistes de la recherche sur les enquêtes a débuté il y a environ 25 ans. Comme beaucoup d'innovations, cette collaboration a eu de nombreux précurseurs et semble émaner de plusieurs sources indépendantes. L'une des plus anciennes, sinon la plus ancienne, a été le colloque sur les problèmes de collecte et d'interprétation de données de

Durant mes études supérieures de deuxième et troisième cycles, j'ai été profondément marqué par le commentaire de Gordon Allport soutenant que le meilleur moyen de découvrir quelque chose consiste à poser une question directe. Plus tard, quand je me suis lancé dans des travaux de recherche sur les problèmes méthodologiques que posent les enquêtes par sondage portant sur des populations humaines, la sagesse de sa remarque m'a semblé encore plus évidente. J'en ai même tiré la Loi pour les questionnaires de Bradburn : « Demandez la chose que vous voulez savoir et non une autre ».

Le problème est qu'il est fort difficile d'appliquer cette loi en pratique pour plusieurs raisons. Premièrement, elle suppose que nous savons ce que nous voulons savoir. Or, souvent, au moment où nous commençons à élaborer un questionnaire, nous n'en sommes pas certains et nous utilisons le processus d'élaboration du questionnaire de façon itérative pour préciser notre idée. Tant que nous ne comprenons pas clairement à propos de quoi nous voulons obtenir des renseignements, il y a peu d'espoir que nous puissions poser des questions sensées.

Deuxièmement, même si nous savons ce que nous voulons savoir, nous devons comprendre comment les gens répondent aux questions. Les complexités de la communication humaine rendent difficile la création d'un instrument unique, normalisé, qui nous permet de poser nos questions de sorte que les répondants les comprennent de la façon dont nous souhaitons qu'elles le soient et que nous comprenions leurs réponses de la façon dont ils le souhaitent. Selon Belson (1964), qui a étudié en profondeur le problème de la compréhension des questions par les répondants, même avec le questionnaire le mieux conçu, moins de la moitié de l'échantillon comprendra les questions de la façon dont le chercheur a l'intention qu'elles soient comprises. Il ne présente pas de données sur la

MEMBRES DU COMITÉ DE SÉLECTION DE L'ARTICLE WASKBERG (2004-2005)

David R. Bellhouse, (Président), *University of Western, Ontario*
 Gordon Brackstone, *Statistique Canada, Ontario*
 Wayne Fuller, *Iowa State University*
 Sharon Lohr, *Arizona State University*

Présidents précédents :

Graham Kalton (1999 - 2001)
 Chris Skinner (2001 - 2002)
 David A. Binder (2002 - 2003)
 J. Michael Brick (2003 - 2004)

Série Waksberg d'articles sollicités

Le comité de rédaction de *Techniques d'enquête* a décidé de publier une série d'articles annuels sollicités en l'honneur de Joseph Waksberg, pour souligner sa contribution importante à la méthodologie d'enquête. Chaque année nous inviterons un spécialiste renommé de la recherche en sondages à rédiger un article consacré à la rétrospective et à l'examen de la situation courante d'un domaine important de la méthodologie d'enquête. L'auteur reçoit un prix monétaire grâce à une subvention offerte par Westat en reconnaissance de la contribution de Joe Waksberg durant les nombreuses années où il a travaillé pour l'entreprise. L'*American Statistical Association* est chargée de la gestion financière et administrative de la subvention. L'auteur de l'article est choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*.

L'auteur de l'article Waksberg est annoncé à la Joint Statistical Meeting annuelle pendant la session de l'adresse présidentielle et des prix de l'*American Statistical Association*. Dans cette session, on félicite les récipiendaires de prix tels que ceux de section, de chapitre, d'excellence en éducation continue et d'autres prix particuliers conjointement. En particulier, on y souligne le prix Waksberg pour contribution remarquable à la théorie et pratique de la méthodologie d'enquête. Enfin, le gagnant du prix Waksberg paraît dans le livret du programme des primes.

Précédents gagnants du prix Waksberg :

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)

Nominations:

Les nominations d'individus à considérés comme auteurs ou les suggestions pour des sujets devraient être envoyés au plus tard le 3 décembre 2004 au président du comité, David Bellhouse par courrier électronique : bellhouse@stats.uwo.ca ou par télécopieur (519) 661-3813.

ARTICLE SOLlicitÉ WAKSBERG 2004

Auteur: Norman M. Bradburn

Norman Bradburn est le professeur émérite distingué du service Tiffany et Margaret Blake de l'Université de Chicago. Il a passé la plus grande partie de sa carrière en tant que méthodologiste au National Opinion Research Center à l'Université de Chicago où il est présentement Fellow sénior. Ses travaux de recherche se sont concentrés sur l'étude des erreurs non dues à l'échantillonnage dans les enquêtes avec un intérêt particulier pour les aspects cognitifs du processus de question/réponse dans les enquêtes.

Hidiroglou et Patak examinent les propriétés d'un certain nombre d'estimateurs régionaux. Ils classent les estimateurs selon deux types, Horvitz-Thompson et Hájek, et selon le niveau de détail des renseignements auxiliaires requis. Ils examinent les propriétés conditionnelles et inconditionnelles des estimateurs sur le plan analytique ainsi qu'au moyen d'une étude en simulation. Ils concluent que les estimateurs de type Hájek offrent les meilleures propriétés conditionnelles, en ce qui concerne tant le biais que la couverture, mais que ces estimateurs n'ont pas de propriété additive et que leurs poids dépendent du domaine.

Dans leur article, Sverchkov et Pfeffermann abordent une prévision des totaux de population finie en utilisant un modèle de variable pour les unités à l'extérieur de l'échantillon (la répartition du complément de l'échantillon) et éventuellement de certaines covariables. Ils décrivent d'abord la répartition de l'échantillon et la répartition du complément de l'échantillon, puis élaborent une estimation semi-paramétrique du modèle du complément de l'échantillon. Ils proposent une procédure de rééchantillonnage pour l'estimation de l'erreur quadratique moyenne. Ils illustrent la méthode d'exemples et la comparaison à d'autres méthodes dans une étude en simulation.

L'article de Grilli et Pratesi considère le problème de l'estimation des paramètres de modèles ordinaux et binaires à plusieurs niveaux pour des plans de sondage informatifs. Les auteurs étendent la méthode du pseudo maximum de vraisemblance pour faire face à ce problème. Cette méthode utilise l'inverse des probabilités d'inclusion à chaque degré pour pondérer le logarithme de la fonction de vraisemblance. Les propriétés de l'estimateur ainsi obtenu sont testées dans une étude par simulations.

La méthode du bootstrap est également utilisée pour obtenir un estimateur de variance. Rowe et Nguyen examinent l'analyse longitudinale en utilisant des données tirées d'une enquête par panel chevauchant, soit l'Enquête sur la population active du Canada. On peut utiliser des panels longitudinaux successifs de six mois pour produire des estimations portant sur des cohortes de personnes au fil du temps, à la condition de pouvoir identifier les membres des cohortes dans chaque panel. Ils élaborent une fonction de vraisemblance pour les données longitudinales observées dans chaque fenêtre de six mois et ils montrent comment on peut s'en servir pour obtenir des estimations des paramètres d'intérêt. Ils donnent ensuite un exemple de cette approche de l'estimation des données observées.

Enfin, dans un article qui a un certain rapport avec celui de Bradburn, Callens et Croux examinent les niveaux individuel et municipal des prédicteurs de la prise de contact et de la coopération dans l'enquête belge sur la fécondité et la famille, en utilisant des modèles de régression logistique multivariés. Ils examinent certains modèles de contact et de coopération fondés sur des théories sociales qui laissent supposer que différents indicateurs jouent un rôle important, puis ils ajustent les modèles en utilisant des données tirées de l'enquête. Leurs constatations qualitatives, particulièrement en ce qui a trait aux indicateurs du statut socioéconomique (SSE), semblent aller à l'encontre des résultats d'études similaires publiées. Dans leur étude, ils concluent que le SSE est positivement associé à la coopération. Ils offrent certaines explications possibles des résultats observés.

Dans ce numéro

Le présent numéro de *Techniques d'enquête* comprend la quatrième de la série annuelle de communications sollicitées en l'honneur de Joseph Waksberg. Une brève description de la série et une courte biographie de Joseph Waksberg ont paru dans le numéro de juin 2001 de la revue. Je tiens à remercier les membres du Comité de sélection de l'article Waksberg d'avoir choisi Norman Bradburn comme auteur de la communication sollicitée cette année.

Dans son article, intitulé « Comprendre le processus de question et réponse », Norman Bradburn fait l'historique de la conceptualisation du processus d'enquête au cours des 20 dernières années selon laquelle des concepts linguistiques et de psychologie sociale et cognitive ont été appliqués pour nous permettre de mieux comprendre le processus; les approches ainsi que les outils cognitifs ont été adaptés aux fins de la rédaction des questionnaires d'enquête. Il présente un modèle conceptuel de l'interview d'enquête et examine divers processus cognitifs dans la réponse aux enquêtes tels que la compréhension, l'extrapolation, la formulation de la réponse et le taux de réponse. Dans sa conclusion, il expose brièvement les défis et les priorités en ce qui concerne d'autres recherches dans ce domaine.

Dans Dennett et Rao, les auteurs présentent une approche de linéarisation de Taylor pour l'obtention d'estimateurs de la variance qui est plus facile à appliquer que l'approche de linéarisation de Taylor habituelle. La nouvelle méthode fournit un estimateur de variance unique et est applicable dans de nombreuses situations et à de nombreux estimateurs. Les auteurs illustrent la méthode pour les estimateurs de calage, les équations d'estimation et l'échantillonnage à deux phases. Pour les estimateurs de calage, le poids calé est saisi automatiquement dans la formule de la variance, justifiant la pratique courante. Phil Kott, Babubhai Shah et Chris Skinner fournissent des commentaires sur cette étude.

Isaki, Tay et Fuller proposent une nouvelle méthode de pondération des ménages pour le questionnaire détaillé du recensement de 2000 aux États-Unis, utilisant une programmation quadratique pour garantir que les sommes pondérées des caractéristiques des ménages et des particuliers correspondent aux totaux de contrôle obtenus d'après le questionnaire abrégé du recensement ou de l'étude d'évaluation de l'exactitude de la couverture (ACE pour Accuracy and Coverage Evaluation). Les poids sont ensuite arrondis à des nombres entiers. Ces auteurs proposent une procédure jackknife d'estimation qui intègre les effets de l'arrondissement ainsi que les valeurs de contrôle aléatoires provenant de l'ACE. Ils comparent les résultats des méthodes de pondération proposées à ceux des méthodes de pondération de 1990 en utilisant les données du Recensement de 1990.

Les propriétés théoriques de l'estimateur par pondération à l'intérieur de cellules sont étudiées dans l'article de da Silva et Opsomer. Contrairement à de nombreuses autres études sur le sujet, où on considère un modèle de réponse dans lequel les unités de la population sont homogènes à l'intérieur des cellules, il n'est pas nécessaire de spécifier correctement le modèle de réponse. Il est cependant nécessaire de connaître une variable auxiliaire qui est corrélée avec la probabilité de réponse. L'approche proposée peut donc être vue comme étant non-paramétrique. Une étude par simulations explore les propriétés de l'estimateur étudié sous divers scénarios. Les auteurs fournissent également quelques recommandations sur la taille et le nombre de cellules de pondération.

Brick, Kalton et Kim traitent de l'estimation de la variance en présence d'imputation hot-deck à l'intérieur de cellules d'imputation pour des estimateurs linéaires. La décomposition de Särndal (1992) et un modèle pour la variable d'intérêt sont utilisés pour estimer la variance. L'originalité de l'approche proposée vient du fait qu'on conditionne non seulement sur les unités échantillonnées et répondantes mais également sur les unités sélectionnées lors de l'imputation. L'article traite également de l'estimation pour des domaines et une étude par simulations est effectuée pour évaluer la méthode proposée quand certaines hypothèses du modèle ne tiennent pas.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 30, numéro 1, juin 2004

TABLE DES MATIÈRES

Dans ce numéro	1
Article Sollicite Waksberg	
NORMAN M. BRADBURN	5
Article de discussion	
ABDELLATIF DEMNATI et J.N.K. RAO	17
Estimateurs de variance par linéarisation pour des données d'enquête	17
Commentaires :	
PHILLIP S. KOTT	28
BABUBHAI V. SHAH	31
CHRIS SKINNER	32
Réponse des auteurs	35
Articles Réguliers	
CARY T. ISAKI, JULIE H. TSAY et WAYNE A. FULLER	39
Pondération de données d'échantillon reposant sur des contrôles indépendants	39
D. NASCIMENTO DA SILVA et JEAN D. OPSOMER	51
Propriétés de l'estimateur à cellules de pondération sous un mécanisme de réponse non paramétrique	51
J. MICHAEL BRICK, GRAHAM KALTON and JAE KWANG KIM	63
Estimation de variance pour l'imputation hot deck à l'aide d'un modèle	63
MICHAEL A. HIDIROGLOU et ZDENEK PATAK	73
Estimation par domaine par la régression linéaire	73
MICHAEL SYVERCHKOV et DANNY PFEFFERMAN	87
Prévision des totaux de population finie basée sur la distribution échantillonnale	87
LEONARDO GRILLI et MONICA PRATESI	103
Estimation pondérée dans le cadre de modèles multivariés ordinaires et binaires sous un plan d'échantillonnage informatif	103
GEOFF ROWE et HUAN NGUYEN	115
Analyse longitudinale des données de l'Enquête sur la population active	115
MARC CALLENS et CHRISTOPHE CROUX	127
Prise de contact et coopération dans l'Enquête belge sur la fécondité et la famille	127

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

D.A. Binder

G.J.C. Hole

C. Patrick

R. Platek (Ancien président)

M.P. Singh

D. Royce

D. Roy

E. Rancourt (Gestionnaire de la production)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*

D.A. Binder, *Statistique Canada*

J.M. Brick, *Westat, Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

M.A. Hidiroglou, *Statistique Canada*

G. Kalton, *Westat, Inc.*

P. Kott, *National Agricultural Statistics Service*

J. Kovar, *Statistique Canada*

G. Lahiri, *JPSM, University of Maryland*

G. Nathan, *Hebrew University, Israel*

D. Norris, *Statistique Canada*

D. Pfeffermann, *Hebrew University*

Rédacteurs adjoints J.-F. Beaumont, P. Dick, H. Mantel et W. Yung, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, Dr. M.P. Singh, singhmp@statcan.ca (Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Pré Tunney, Ottawa, Ontario), Canada K1A 0T6. Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de *Techniques d'enquête* (n° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 30 \$ CA (15 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commander par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Juillet 2004

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

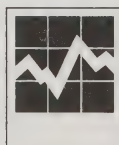
© Ministre de l'Industrie, 2004

Publication autorisée par le ministre
responsable de Statistique Canada

JUN 2004 • VOLUME 30 • NUMÉRO 1

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 1

•

VOLUME 30

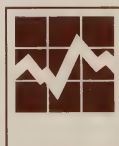
•

JUIN 2004

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE



12-001



SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2004

•

VOLUME 30

•

NUMBER 2



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2004 • VOLUME 30 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2005

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.



January 2005

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder
G.J.C. Hole
C. Patrick
R. Platek (Past Chairman)

E. Rancourt (Production Manager)
D. Roy
D. Royce
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat, Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidirolou, *Office for National Statistics*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
J. Kovar, *Statistics Canada*
P. Lahiri, *JPSM, University of Maryland*
G. Nathan, *Hebrew University*
D. Norris, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R. Valliant, *JPSM, University of Michigan*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *Iowa State University*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, Dr. M.P. Singh, singhmp@statcan.ca (Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$30 (\$15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

SURVEY METHODOLOGY
A journal Published by Statistics Canada
Volume 30, Number 2, December 2004

CONTENTS

In This Issue	125
Discussion Paper	
PAUL P. BIEMER	
An Analysis of Classification Error for the Revised Current Population Survey Employment Questions	127
Comment:	
JEROEN K. VERMUNT	141
STEPHEN M. MILLER and ANNE E. POLIVKA	145
CLYDE TUCKER	151
Response from the author	154
Regular Papers	
PATRICIA GUNNING and JANE M. HORGAN	
A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations	159
DAN HEDLIN and SUOJIN WANG	
Feeding Back Information on Ineligibility from Sample Surveys to the Frame	167
WALTER MUDRYK and HANSHENG XIE	
Application of Quality Control in ICR Data Capture: 2001 Canadian Census of Agriculture	175
INHO PARK and HYUNSHIK LEE	
Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling	183
JEAN-FRANÇOIS BEAUMONT and ASMA ALAVI	
Robust Generalized Regression Estimation	195
HUI ZHENG and RODERICK J.A. LITTLE	
Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples	209
FAMING LIANG and ANTHONY YUNG CHEUNG KUK	
A Finite Population Estimation Study with Bayesian Neural Networks	219
JEROME J. REITER	
Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation	235
Acknowledgements	243
Erratum	244

In This Issue

This issue of *Survey Methodology* opens with a discussed paper by Paul Biemer. He provides evidence of reduced accuracy due to the redesign of employment questions in the Current Population Survey (CPS). This is an extension of the previous study by Biemer and Bushery (2000). In the current paper, the author attempts to trace the source of the error through extended analysis of the CPS data before and after the redesign. A new approach, using Markov Latent Class Analysis, is presented. This work aims at providing guidance for further investigation into the root causes of the errors in the collection of labour force data in the CPS. Discussions of this paper are provided by Jeroen Vermunt, Stephen Miller and Anne Polivka, and Clyde Tucker.

In their paper, Gunning and Horgan propose a new algorithm for the construction of stratum boundaries in skewed populations. Their algorithm uses an auxiliary variable and achieves equal coefficients of variation for this auxiliary variable in each stratum. The method is based on the assumption that the auxiliary variable is uniformly distributed. One advantage of the method is that it is very easy to apply in practice. In an empirical study, the authors show that the proposed algorithm compares favourably with the cumulative root frequency method of Dalenius and Hodges (1957) and to the Lavallée and Hidioglou (1988) algorithm.

Hedlin and Wang consider the problem of bias coming from feeding back information from sample surveys to frames. They investigate the bias incurred by updating deaths on a frame that is used for future occasions of the same survey. They quantify this bias and develop an unbiased estimator for this situation. The theoretical results presented in the paper are illustrated through a simulation study.

In their paper, Mudryk and Xie present the Quality Assurance (QA) and Quality Control (QC) aspects of the Intelligent Character Recognition operation of the 2001 Canadian Census of Agriculture. They show how an effective QA and QC plan was developed to ensure the highest quality data from the data capture operation of the Census. Results from an analysis of the Average Outgoing Quality of the data indicate the importance of a QA/QC plan.

In Park and Lee, the design effects for the weighted mean and total estimators are investigated for complex surveys. In particular, they decompose the design effect for the weighted mean and total estimators under a two-stage design. Given this decomposition, they illustrate several common misconceptions about the design effects for the weighted mean and total estimators through several examples using commonly used designs.

In their paper, Beaumont and Alavi investigate a robust generalized regression estimator. They look at alternatives to the optimal Best Linear Unbiased (BLU) estimator that are robust to design ignorability and/or model misspecification. In the situation where the design ignorability assumption may not hold, they propose a least squares estimator that is obtained by shrinking the design weights to their mean. To deal with model misspecification, they propose a weighted generalized M -estimator to reduce the influence of units with large weighted population residuals. Their theoretical results are illustrated with a simulation study.

Zheng and Little propose a non-parametric model-based alternative to Horvitz-Thompson estimation of a total in the case of two-stage sampling with pps sampling at the first stage. This is an extension of their earlier work in which an outcome variable y_i is modeled as a smooth function of the inclusion probability π_i . They show how to fit the model and estimate the total using a penalized spline, and also develop alternative variance estimation procedures. Simulations are used to compare the proposed method to the Horvitz-Thompson estimator and to a model-assisted estimator.

Liang and Kuk consider an alternative to the standard approach for regression estimation in a finite population. Instead of the usual linear model they use an arbitrary smooth function to allow for a non-linear regression, and then they apply Bayesian neural networks to the problem. The advantage of the neural network approach is that the problem of model misspecification is avoided. Liang and Kuk place a prior on each network connection instead of on the number of hidden units as is usually done. This permits a unified approach to the selection of the network structure and the selection of the auxiliary variables. Finally, they handle outliers by introducing a heavy tail distribution to model the disturbances of the data.

In the last paper of this issue, Reiter uses multiple imputation to handle simultaneously both missing data and disclosure limitation. The basic idea is to fill in the missing data first to generate m completed datasets and then replace sensitive or identifying values in each completed dataset with r imputed values. Then, the author develops new combining rules for obtaining valid inferences from such multiply-imputed datasets. These rules take into account both sources of variability in the point estimators.

Finally, the Editorial Board met this past summer at the Joint Statistical Meetings in Toronto. A suggestion was made at that meeting to have a Short Communications section in the journal. These would be shorter papers, typically around four Survey Methodology pages. Possible topics of short communications would include presentation of new ideas without the full development of a regular paper, brief reports of empirical work, and discussions or supplements to other papers published in the journal. All short communications would be refereed, although the reviewing process may be streamlined. I hope that this new format will be attractive to many authors, and look forward to receiving your submissions.

M.P. Singh

An Analysis of Classification Error for the Revised Current Population Survey Employment Questions

PAUL P. BIEMER¹

ABSTRACT

The reduced accuracy of the revised classification of unemployed persons in the Current Population Survey (CPS) was documented in Biemer and Bushery (2000). In this paper, we provide additional evidence of this anomaly and attempt to trace the source of the error through extended analysis of the CPS data before and after the redesign. The paper presents an novel approach decomposing the error in a complex classification process, such as the CPS labor force status classification, using Markov Latent Class Analysis (MLCA). To identify the cause of the apparent reduction in unemployed classification accuracy, we identify the key question components that determine the classifications and estimate the contribution of each of these question components to the total error in the classification process. This work provides guidance for further investigation into the root causes of the errors in the collection of labor force data in the CPS possibly through cognitive laboratory and/or field experiments.

KEY WORDS: Survey redesign; Measurement error; Latent class analysis; Unemployment rate; Specification error.

1. INTRODUCTION

The Current Population Survey (CPS) is a monthly survey of approximately 60,000 households conducted by the U.S. Bureau of the Census for the Bureau of Labor Statistics (BLS). The primary purpose of the survey is to provide estimates of employment, unemployment, and other characteristics of the general U.S. labor force population. Estimates of the size, composition, and dynamic characteristics of the labor force are published each month by BLS and comprise one of the Nation's key economic indicators.

In January 1994, a revised questionnaire was introduced in the CPS to address the recommendations by the Levitan Commission in the late 1970s to convert the mode of interview for the CPS from paper and pencil questionnaire to computer-assisted interviewing methods, to clarify some of the questions on employment, as well as for a number of other reasons described in Rothgeb (1994). The overall objective of the redesign was to improve the quality of the data collected in the CPS. The CPS questionnaire had remained essentially unchanged since the last major revision in 1967.

The revised CPS questionnaire was introduced after considerable research and testing that began in the mid-1980s. The purpose of the testing was to evaluate the quality and operational feasibility of various redesign options including moving the CPS from a paper and pencil questionnaire format to computer assisted interviewing. During these years of testing, more than 100,000 persons were interviewed in the various studies that were conducted (Rothgeb 1994). The CPS redesign research program

culminated in a large national study (referred to in the literature as the CATI/CAPI Overlap or CCO Field Test) that was conducted in 1993. The key component of this test consisted of a computer assisted survey of approximately 12,000 households implementing revised CPS interviewing procedures and the revised questionnaire. This survey, referred to in this report as the Parallel Survey, was conducted from July 1992 to December 1993 concurrently with the ongoing CPS survey which used the original questionnaire. This type of split panel design makes it possible to estimate the effect of the redesign changes on the CPS labor force estimates.

A number of papers and reports were published documenting the findings from the CCO Field Test (Cohany, Polivka and Rothgeb 1994; Rothgeb 1994; Polivka 1994; Kostanich and Cahoon 1994; Miller 1994; Thompson 1994; Dipbo, Polivka, Creighton, Kostanich and Rothgeb 1994). One key finding from this research was that the Parallel Survey unemployment rate and the labor force participation rate were higher than in the CPS. The higher unemployment and labor force participation rates associated with the revised questionnaire were explained primarily by changes in the definition of employment. The revised questionnaire has a broader approach to both work and job search activities, which would tend to classify more persons as "in the labor force" and, thus, more persons who are not working as unemployed rather than out of the labor force (see, for example, Polivka 1994 and Rothgeb 1994).

The increase in the unemployment rate due to the new design was originally estimated at about one-half percentage point. However, further analysis of the Parallel Survey data

¹ Paul P. Biemer, 3040 Cornwallis Road, PO Box 12194 Research Triangle Park, NC 27709-2194, U.S.A.

called that estimate into question and subsequently a report was release estimating the increase to be less than one-tenth percentage point (Polivka and Miller 1994). The concerns raised in the subsequent reports regarding the utility of the Parallel Survey data for assessing the effect of the redesign are discussed further below and will be considered in our analysis of these data.

An independent analysis conducted by Biemer and Bushery (2000) revealed an anomaly in the revised CPS labor force data that had not been detected by any of the previous research on the CPS redesign. Using a Markov latent class analysis (MLCA) approach, Biemer and Bushery compared the accuracy of labor force classifications under the original and revised designs by estimating and comparing the error rates using the 1993 CPS data and the 1995 and 1996 CPS data. They defined labor force classification accuracy as the probability that a person who is truly in some labor force category, say category a , is classified as being in a by the CPS; *i.e.*, $\Pr(\text{classified in } a \mid \text{truly in } a)$. For example, the classification accuracy for unemployment is the probability a person who is truly unemployed, according to the CPS definition, is correctly classified as unemployed by the CPS classification rules.

In Table 2 of their paper, Biemer and Bushery report that the classification accuracy for unemployment dropped by 5.7 percentage points, from approximately 81.8 percent (*s.e.* = 0.90) in 1993 to 76.1 (*s.e.* = 1.2) in 1995 and 74.4 percent (*s.e.* = 1.2) in 1996. These results suggest that the redesigned CPS misclassifies the true unemployed at a higher rate than the old CPS design. The authors first considered that this result could be an artifact of the MLCA methodology. As shown below, MLCA does not require a true or "gold standard" measurement of employment to estimate classification error. Rather the method relies a model describing the true month to month changes in employment status and as well as for the process of classifying individuals into labor force categories. It is possible that labor force transitions that deviate from the model specification could be regarded as misclassifications in the estimation process.

To check the validity of the MLCA results, the authors conducted a series of analyses using traditional estimation approaches, analysis of the error by population groups, comparisons of the error estimates to other published estimates, and simulations to assess the effect of model failure on the results. As an example, there is evidence that the test-retest reliability of the unemployment category decreased after the redesign. Prior to the redesign, the index of inconsistency (The index of inconsistency is a measure of unreliability traditionally used at the Census Bureau. It is equal to $1 - \kappa$ where κ is Cohen's kappa coefficient (Cohen 1960) for the unemployed labor category averaged 30

percent for the period 1992–1993. Following the redesign, the index of inconsistency increased to almost 40 percent for the period 1995–1996. These analyses support their claim that the accuracy of the CPS methodology for classifying unemployed persons declined after the redesign.

In their discussion of the results, the authors speculated that the drop in classification accuracy could indicate a problem with the revised unemployment questions. That is, the revised unemployment questions may be subject to greater classification error and, thus, less classification accuracy. Another possibility they considered is change in the characteristics of the unemployed populations from 1993 to 1995 and 1996. Since the unemployment rate dropped from 1993 to 1996, it is possible that persons who would be more accurately classified by the CPS system left the ranks of the unemployed, leaving persons who would be less accurately classified in the category. This hypothesis could be tested by estimating the accuracy rates for the two methodologies for the same time period. The Parallel Survey offers a means to conduct such an analysis.

The current paper continues the investigation of the reduction in MLCA unemployment classification accuracy rates observed by Biemer and Bushery. The current analysis uses MLCA models very similar to those used by Biemer and Bushery for estimating the classification accuracy for the original and revised versions of the CPS questionnaire. However, the time period considered here is expanded to include the 15 months prior to and following the introduction of the revised questionnaire: a total of 30 contiguous months. In addition, data from the Parallel Survey from the period January 1993 through December 1993 is used to compare the employment accuracy for original and revised questionnaire for the same time period.

Our analysis focuses on a labor force classification variable that is derived from a number of questions on the employment section of the CPS questionnaire. This variable is often referred to as a "recoded" labor force variable since it is determined by mapping a pattern of CPS responses to questions about employment onto particular labor force categories such as employed – at work, employed – not at work, unemployed – looking for work, and so on. Biemer and Bushery used a three-category employment classification variable: employed (EMP), unemployed (UEM), and not in the labor force (NLF). For the present analysis, a four-category variable is used that subdivides the UEM category into unemployed-on layoff (UEM-LAYOFF) and unemployed-looking for work (UEM-LOOKING). This is done as a first step toward isolating the source of the apparent inaccuracy in unemployment classification. However, further decomposition of these categories will be necessary to arrive at the root source of the error as will be shown subsequently.

In section 2 we describe the CPS labor force concepts that are most relevant to our study and the structure of the data sets in the analysis. In section 3 we review the MLCA estimation methodology and models used by Biemer and Bushery in their analysis and describe the application of their methodology for the present purposes. In section 4 we present the results of our analysis and what they suggest regarding the source of the classification error in the new questionnaire. Finally, section 5 provides a summary of the key findings and our conclusions from the study.

2. DATA AND CONCEPTS

2.1 The Data Sets for Our Study

Except for the Parallel Survey, the CPS data in our analysis were downloaded from the National Bureau of Economic Research (NBER) website (www.nber.org). This website contains microdata for the CPS for every month from January 1976 through December 2004. The MLCA approach was applied directly to these microdata without the need for supplementary data or data external to the CPS.

In the preliminary analysis, we investigated the CPS classification accuracy for a six-year period: January 1992 through December 1997. That analysis was aimed at determining whether the anomaly first noted in Biemer and Bushery (2000) is a transient phenomenon affecting only the months immediately following the introduction of the new questionnaire or whether it persisted for some years after the new questionnaire was introduced. If temporary or transient, the anomaly might be related to problems during the phase-in of the new design; for example, interviewer training or issues related to the startup of data collection. However, evidence of a persistent, continuing effect could suggest problems with the survey design; for example, the questionnaire, interviewing procedures, or the recoding algorithm.

By applying MLCA across all months from 1992 through 1997 we determined that, although the magnitude of the reduction in accuracy varies somewhat from month to month, it does indeed persist for all months following the introduction of the revised questionnaire. The results confirmed Biemer and Bushery's conjecture of a systemic effect possibly linked to the new unemployment questions introduced in January 1994.

Due to space considerations, in this paper we present results from a somewhat shorter time frame than considered in the preliminary analysis, *viz.*, the years 1992, 1993, 1994, and 1995. This time period covers two years of the CPS using the original questionnaire and two years using the revised questionnaire. In addition, we will also present some results from an MLCA of the 1993 Parallel Survey data that can be compared with results from the main CPS.

The data sets in our study are quite large. Each estimate of classification error we obtain is based upon all households that were interviewed in the CPS for three consecutive months. Across the four years in our analysis, the total number of households responding for all three months in any three-month period varies from about 37,000 to more than 40,000. For the 1993 Parallel Survey, the number of households satisfying this criterion is approximately 10,000. The estimates we produce are appropriately weighted for probabilities of selection and other post-survey adjustments and, therefore, reflect the response probabilities of the published CPS estimates. Weights were constructed by taking an average weight across the three consecutive months that were combined to form a longitudinal record for the analysis (unweighted analyses were also conducted and the results were very similar to the weighted analysis. This suggests the choice of weights has little effect on the study outcomes).

Because of a problem in the identification variables required for linking households for the months June 1995 through December 1995, it was not possible to include these months in our analysis. Further, since our conclusions would not change by including data from the 1996 or later years of the CPS, we confine our analysis to 15 months prior and 15 months following the introduction of the revised questionnaire. Thus, for most of the analysis to follow, we will provide averages of estimates from August 1992 through December 1993 for the original questionnaire and from January 1994 through May 1995 for the revised questionnaire (note that since our estimates are based upon a moving average of three consecutive months, seasonal variations in the labor rates and transitions probabilities are accounted for in the estimates of classification error).

2.2 Labor Force Concepts

The revised CPS questionnaire was introduced in 1994 to improve the overall quality of labor market information through extensive question changes and through the use of computer technology in the data collection. In the following, we describe a few concepts that were affected by the questionnaire redesign and that are relevant for the current analysis.

Employed. The labor force questions in the original questionnaire began with the question "What were you doing most of LAST WEEK (working, keeping house, going to school, or something else)?" Interviewers were allowed to modify the parenthetical part of this question according to the age of the respondent. In some cases, the word "work" or "working" was not part of the question. As an example, if the respondent looked of student-age, the interviewer was allowed to leave out the word "working." The revised questionnaire replaced this question with two

questions: "Does anyone in this household have a business or a farm?" and "LAST WEEK, did you do ANY work for (either) pay (or profit)?" where the parenthetical parts of the question are read if anyone in the response to the first question is "yes." Further, additional questions were added to clarify whether earnings or profits were received from the family business or farm. Thus, the revised questionnaire concept of employment appears to be somewhat broader and better defined than the original questionnaire concept.

Unemployed. The definition of unemployment was slightly modified in the revised questionnaire. In the original questionnaire, persons waiting for a new job to start were classified as unemployed. Under the revised questionnaire definition, a person is unemployed only if all of the following are true: (1) without a job, (2) actively seeking work or on layoff from a job and expecting recall within the next six months, and (3) currently available to take a job (except for a possible temporary illness).

On Layoff. Persons on layoff are defined as persons separated from a job and who are awaiting a recall to return to that job. The original questionnaire did not consider or collect information on the expectation of recall. This was problematic because to most people, the term "layoff" could mean permanent termination from the job rather than the temporary loss of work economists are trying to measure.

Job Search Methods. To be counted as unemployed and looking for work, a person must have engaged in an active job search during the four weeks prior to the survey. The revised questionnaire includes a somewhat broader question about job search methods with expanded and restructured response categories to allow interviewers to more easily record and distinguish between active and passive job search activities. In addition, it provides additional followup questions for those who respond "nothing" or "don't know."

Reference Week. While the original questionnaire referred to LAST WEEK, the reference period was never

explicitly defined. The revised questionnaire provides specific dates of the reference week.

We will refer to these changes later in the report when we discuss the differences in the classification error and specification error between the revised and original questionnaires.

As previously noted, Biemer and Bushery focused on a three-category labor force recoded variable with categories: employed (EMP), unemployed (UEM), and not in the labor force (NLF). For the present analysis, we used an expanded recoded variable also available on the CPS public use data files. This variable divides the UEM category into two categories corresponding to persons on layoff (LAYOFF) and persons looking for work (LOOKING). The seven-category variable also divides the EMP and NLF categories into subcategories; however, this level of detail in the EMP and NLF categories is not needed in our analysis. Thus, the seven-category variable will be collapsed to a four-category variable corresponding to EMP, UEM-LOOKING, UEM-LAYOFF, and NLF. The correspondence between the three- and four- category variables is shown in Figure 1.

3. LATENT CLASS MODELS FOR CPS CLASSIFICATION ERROR

Markov latent class models were first proposed by Wiggins (1973) and refined by Poulsen (1982). Van de Pol and de Leeuw (1986) established conditions under which the model is identifiable and gave other conditions of estimability of the model parameters. In this section we describe the basic model proposed by Biemer and Bushery (2000) and its extensions for application in the current analysis.

Let the CPS target population be divided into L groups (such as age, race, or sex groups) and let the variable G be the label for group membership. For example, $G_i = 1$ if the

Original Seven-Variable Category Old Questionnaire	New Questionnaire	Four-Category Analysis Variable	Three- Category Analysis Variable
1. Working—at work	1. Employed—at work	1. EMP	1. EMP
2. With job—not at work	2. Employed—absent		
3. Unemployed—on layoff ¹	3. Unemployed—on layoff	2. UEM—LAYOFF	2. EM
4. Unemployed—looking for work ¹	4. Unemployed—looking	3. UEM—LOOKING	
5. Working without pay (less than 15 hours in a family farm or business) or temporarily absent from a without pay job	5. Retired—not in labor force	4. NLF	3. NLF
6. Unavailable to take a job if one had been offered	6. Disabled—not in labor force		
7. Not in the labor force	7. Other—not in labor force		

¹ Note: In the original questionnaire, categories 3 and 4 are reversed compared to corresponding categories in the revised questionnaire.

Figure 1. Association of the Seven-Category Employment Recode Variable with the Three- and Four-Category Variables Used in the Analysis

i^{th} population member is in group 1, $G_i = 2$ for group 2 and so on. Let X_{gi} , Y_{gi} , and Z_{gi} denote the true labor force classifications for the i^{th} person in group $G = g$ (for $g = 1, \dots, L$ and $i = 1, \dots, n_g$) where X_{gi} is defined as

$$X_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is employed in time period 1} \\ 2 & \text{if person } (g, i) \text{ is unemployed -} \\ & \text{on layoff in time period 1} \\ 3 & \text{if person } (g, i) \text{ is unemployed -} \\ & \text{looking in time period 1} \\ 4 & \text{if person } (g, i) \text{ is not in the labor force} \\ & \text{in time period 1} \end{cases}$$

with analogous definitions for Y_{gi} and Z_{gi} for periods 2 and 3 respectively. Consistent with the conventions of the LCA literature, we will drop the subscripts from the variables to simplify the notation.

Let $\pi_{x,y,z|g}$ denote $\Pr(X = x, Y = y, Z = z | G = g)$, let $\pi_{y|g,x}$ denote $\Pr(Y = y | X = x, G = g)$ and let $\pi_{z|g,y,x}$ denote $\Pr(Z = z | Y = y, X = x, G = g)$. Then, the probability that an individual in group g has labor status x in period 1, y in period 2, and z in period 3 is $\pi_{xyz|g}$ which may be written as

$$\pi_{xyz|g} = \pi_{x|g} \pi_{y|gx} \pi_{z|gxy}. \quad (1)$$

Finally, under the first order Markov assumption, which is a necessary condition for model identifiability (see Van de Pol and de Leeuw 1986), we assume

$$\pi_{z|gxy} = \pi_{z|gy} \quad (2)$$

i.e., at period 3, the true status of an individual does not depend on the period 1 status, once the period 2 status is known. An alternate interpretation is that the current status, given the prior period's status, does not depend upon the prior period's transition.

Now, consider the observed labor force classifications from the CPS denoted by A_{gi} , B_{gi} , and C_{gi} for periods 1, 2, and 3, respectively, where

$$A_{gi} = \begin{cases} 1 & \text{if person } (g, i) \text{ is classified as EMP in time} \\ & \text{period 1} \\ 2 & \text{if person } (g, i) \text{ is classified as UEM -} \\ & \text{LAYOFF in time period 1} \\ 3 & \text{if person } (g, i) \text{ is classified as UEM -} \\ & \text{LOOKING in time period 1} \\ 4 & \text{if person } (g, i) \text{ is classified as NLF in} \\ & \text{time period 1} \end{cases}$$

with analogous definitions for the response indicators, B_{gi} , and C_{gi} for periods 2 and 3, respectively. Using an extension of the notation established above, we denote the response probabilities in each of these classifications as $\pi_{a|gx} = \Pr(A = a | X = x)$, with analogous definitions for $\pi_{b|gy}$ and $\pi_{c|gz}$. Thus, $\pi_{a=1|g,x=2}$ is the probability that the CPS classifies a person in group g as employed ($A = 1$) when the true status is unemployed – on layoff ($X = 2$). Likewise, $\pi_{a=2|g,x=2}$ is the probability that the CPS correctly classifies a person in group g as unemployed – on layoff.

Finally, we assume

$$\pi_{a,b,c|g,x,y,z} = \pi_{a|gx} \pi_{b|gy} \pi_{c|gz} \quad (3)$$

or that classification error in the observed labor force status is independent across the three months.

The CPS labor force classifications for each month of a three consecutive month interval are the outcome variables in our analysis. Let A , B , and C denote the observed classifications and let X , Y , and Z denote the (unobserved) true classifications for Month 1, Month 2, and Month 3, respectively. Let G denote some grouping (or stratification) variable to be defined later in the analysis. Under these assumptions, we can write the probability for classifying a CPS sample member in cell (g, a, b, c) of the $GABC$ table as follows:

$$\pi_{gabc} = \sum_{x,y,z} \pi_g \pi_{x|g} \pi_{y|gx} \pi_{z|gy} \pi_{a|gx} \pi_{b|gy} \pi_{c|gz}. \quad (4)$$

Extensions to more than one grouping variable are straightforward.

Under multinomial sampling, the likelihood function for the $GABC$ table is

$$\Pr(GABC) = C \prod_{g,a,b,c} \pi_{gabc}^{n_{gabc}} \quad (5)$$

where C is the multinomial constant and Π denotes the product of the terms over the subscripts g , a , b , and c . Under the assumptions made previously, the model parameters are estimable using maximum likelihood estimation methods. Van de Pol and de Leeuw (1986) provide the formula for applying the E-M algorithm to estimate the parameters of this model and describe the conditions for their estimability. The ℓ EM software (Vermunt 1997) was used to fit the MLCA models.

In their investigations of the validity of MLCA estimates for analyzing CPS labor force classification error, Biemer and Bushery analyzed CPS data collect during the first quarter of each of three years – 1993, 1995, and 1996. They also conducted several types of analysis using the CPS unreconciled reinterview data for the same time period. The reinterview analysis provided another approach for

estimating CPS classification error as well as evidence of the validity of the MLCA approach. Their evaluation of MLCA validity considered five criteria: (1) model diagnostics, (2) model goodness of fit across years of CPS, (3) agreement between the model and test-retest estimates of response probabilities, (4) agreement between the model and test-retest estimates of inconsistency, and (5) plausibility of the patterns of classification error. The MLCA method performed well in all five test. For example, the same model provided the best fit of the data for each year analyzed, there was good agreement between the latent class estimates of reliability and those derived from traditional test-retest methodology; and the estimated error rates were consistent with those of previous studies – for *e.g.*, Chua and Fuller 1987; Abowd and Zellner 1985; Porterba and Summers 1995; and Sinclair and Gastwirth 1998.

Ostensibly, the Markov assumption seems very unlikely to hold for labor force data. As an example, persons who are unemployed in months 1 and 2 of a consecutive three-month period may not have the same probability of being unemployed in a month 3 as persons who just became unemployed in month 2. The former group could contain more chronically unemployed persons than the group entering unemployment in month 2. Further, the group just entering unemployment in month 2 could contain a higher proportion of people temporarily out of work while changing jobs. Biemer and Bushery considered the consequences for the MLCA estimates of misclassification when the Markov assumption is violated.

Using simulation, Biemer and Bushery found that the bias in MLCA estimates of classification probabilities depends upon the severity of the departures of the CPS data from the Markov assumption. They defined two parameters, λ_1 and λ_2 , which are ratios of conditional probabilities. λ_1 is the ratio of the probability of being employed in period 3 for a person with an (EMP, UEM) pattern for periods 1 and 2, respectively, divided by the probability of being employed in period 3 for a person with a (EMP, EMP) pattern. Similarly, λ_2 is the ratio of the probability being employed in period 3 for a person with an (UEM, UEM) pattern to the probability of being employed in period 3 for a person with a (EMP, UEM) pattern. Note that when $\lambda_1 = \lambda_2 = 1$, the Markov assumption holds exactly and greater departures of λ_1 and λ_2 from 1 correspond to greater departures of the data from the Markov assumption. Biemer and Bushery found that over a fairly wide range of values for λ_1 and λ_2 , the absolute bias in the MLCA estimates of unemployment classification accuracy never exceeded 3 percentage points. For example, in the extreme case of a Markov assumption violation, the expected value of an MLCA estimate of unemployment accuracy would be 77 percent when the true parameter value is 80 percent.

Their results suggest that, for the CPS application, MLCA is fairly robust to failures of the Markov assumption to hold.

Although it is virtually impossible to prove their validity, MLCA error estimates can be quite useful for identifying survey questions that are prone to classification error; *i.e.*, flawed questions. For example, Biemer (2004) and Biemer and Wiesen (2002) demonstrate the utility of MLCA methodology for identifying question problems and classification process deficiencies in large scale surveys. Notwithstanding that the MLCA assumptions may be violated to an unknown extent, its usefulness as a tool for exploring a number of important questionnaire design issues has been well-documented. For the present application, MLCA will be used to develop and test hypotheses regarding the sources of the anomaly reported by Biemer and Bushery for 1994 CPS redesign.

The MLCA model use in the present analysis is essentially the same model selected by Biemer and Bushery for their analysis. To account for population heterogeneity, they considered a number of demographic and other explanatory variables that might be highly correlated with classification error. The best performing variable a proxy or self-response indicator variable denoted by P where

$$P = \begin{cases} 1 & \text{if all three interviews are conducted by self response (SELF)} \\ 2 & \text{if two of the interviews are conducted by self response (MOSTLY SELF)} \\ 3 & \text{if two of the interviews are conducted by proxy response (MOSTLY PROXY)} \\ 4 & \text{if all three interviews are conducted by proxy response (PROXY).} \end{cases}$$

Their empirical findings showed this variable to be strongly related not only to reporting accuracy, but also current employment status and month to month employment transitions. For example, responses for the PROXY group were considerably less accurate than for the SELF group and, further, the PROXY group had somewhat higher unemployment than the SELF group.

The MLCA model also allows transition probabilities to vary by P (referred to as group heterogeneity) as well as by time periods (referred to as non-stationary transitions). In addition, the model assumes that response probabilities $\pi_{a|px}$, $\pi_{b|px}$, and $\pi_{c|px}$ are group-heterogeneous but are equal for all three months in the time interval. This leads to the following model for describing the cell probabilities in the PABC table:

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|p} \pi_{y|px} \pi_{z|py} \pi_{a|px}^{A|PX} \pi_{b|py}^{A|PX} \pi_{c|pz}^{A|PX} \quad (6)$$

where $\pi_{b|py}^{APX} = \Pr(A=b | P=p, X=y)$ with similar definitions for $\pi_{a|px}^{APX}$ and $\pi_{a|pz}^{APX}$. That is, the three sets of response probabilities are equal to π^{APX} .

Note that for the present analysis, interest is focused on the overall response probabilities associated with the revised and original questionnaires and not the variation in error rates across proxy groups. Therefore, our analysis focuses on the overall accuracy of response, *i.e.*, π_{dx}^{Alx} or the mean response probability for the four levels of P combined.

4. COMPARISON OF REVISED AND ORIGINAL QUESTIONNAIRE CLASSIFICATION ERROR PROBABILITIES

4.1 Reduction in UEM Classification Accuracy for the Revised Questionnaire

As mentioned in section 2, the CPS data sets for this analysis are monthly samples from August 1992 through May 1995. Figure 2 shows how this the time interval was divided into 30 overlapping three-month intervals: 15 for the original questionnaire and 15 for the revised questionnaire. The intervals are numbered in the table for later reference. For example, time interval 1 covers the period from August 1992 through October 1992 in which the original questionnaire was in use. Therefore, this time interval can provide one estimate of the response probabilities, π^{Alx} , for the model in (6). Since there are 30 time intervals across the entire 34-month period in our analysis, 30 estimates of π^{Alx} can be formed from these consecutive overlapping time intervals: 15 estimates for the original questionnaire and 15 estimates for the revised questionnaire.

To obtain a more stable estimate of π^{Alx} for each questionnaire, the 15 estimates corresponding to the 15 time periods per questionnaire in Figure 2 were averaged. These estimates are shown in Tables 1 and 2. Since they are based on simple random sampling assumptions, the standard errors in the tables do not account for the unequal weighting

and clustering effects of the CPS. Since the average CPS design effect is about 1.5 for estimates of unemployment, the standard errors in the tables are probably understated by 20 percent or less. This level of bias in the standard errors is inconsequential for the purposes of this paper due to the extremely large sample sizes in the analysis.

Table 1 compares the MLCA estimates of the classification error probabilities for the original and revised questionnaire versions for the three-category labor force classification scheme used by Biemer and Bushery. The first column of the table is the true (or latent) category, the second column is the observed (or CPS) category, and the cell entries are the response probabilities estimated from the MLCA using model (6). For each true class (EMP, UEM, or NLF), the accuracy rate is the cell corresponding to the observed category with the same label. For example, the accuracy of classifying persons who are truly employed is 98.68 percent (for the original questionnaire) and 98.84 percent for the revised questionnaire. Note that this entry corresponds to the cell where both the true category and the observed category are EMP. The other cells for EMP in column 1 are the error rates for EMP. For example, the MLCA estimate of the probability CPS classifies a person as UEM who is truly EMP is 0.42 for the original questionnaire and 0.39 for the revised questionnaire. The other cell entries are interpreted analogously.

Consistent with Biemer and Bushery's findings, the accuracy of the classification of unemployed persons is substantially and highly significantly lower for the revised questionnaire: 79.06 percent versus 73.50 percent, a difference of 5.6 percentage points. Further, the increase in classification error for unemployed persons is due to misclassifications in both the EMP and NLF force categories with slightly more misclassification in the latter category. Our estimates differ slightly from theirs since, as noted earlier, we are analyzing more months of data and using weighted estimates rather than unweighted as in their analysis.

Months Questionnaire	Using	Old	Aug. 1992	Sept. 1992	Oct. 1992	Nov. 1992	Dec. 1992	Jan. 1993	...	Aug. 1993	Sept. 1993	Oct. 1993	Nov. 1993	Dec. 1993
Month Questionnaire	Using	New	Jan. 1994	Feb. 1994	March 1994	Apr. 1994	May 1994	June 1994	...	Jan. 1995	Feb. 1995	March 1995	Apr. 1995	May 1995
Interval														
1 (Old), 16 (New)		X		X	X									
2 (Old), 17 (New)				X	X	X								
3 (Old), 18 (New)					X	X	X							
4 (Old), 19 (New)						X	X	X						
...														
13 (Old), 28 (New)										X	X	X		
14 (Old), 29 (New)											X	X	X	
15 (Old), 30 (New)												X	X	X

[†] The "..." symbol is used in this table to indicate that the pattern established for the preceding months continues for the remaining months.

Figure 2. The 30 Three-Month Time Intervals Analyzed for the Revised and Original Questionnaires

Table 1
Comparison of CPS Labor Force Response Probabilities for the Original and Revised Questionnaires

True Class	Observed Class	Original (1992–1993)	Revised (1994–1995)	Original – Revised Diff	S.E.
EMP	EMP	98.68	98.84	–0.15	0.40
	UEM	0.42	0.39	0.03	0.40
	NLF	0.90	0.78	0.13	0.16
UEM	EMP	8.23	10.57	–2.34*	0.45
	UEM	79.06	73.50	5.56*	0.54
	NLF	12.71	15.93	–3.32*	0.26
NLF	EMP	2.14	1.99	0.15	0.36
	UEM	1.43	1.56	–0.13	0.33
	NLF	96.43	96.45	–0.02	0.18

* Significant at $\alpha = 0.001$.

Table 2
Comparison of Two Unemployed Subcategories for the Original and Revised Questionnaires

True Class	Observed Class	Original (1992–1993)	Revised (1994–1995)	Original – Revised Diff	S.E.
UEM–LAYOFF	EMP	16.32	26.67	–10.35*	0.91
	UEM – Layoff	61.30	55.63	5.66*	1.03
	UEM – Looking	17.61	8.41	9.20*	0.45
	NLF	4.77	9.29	–4.52*	0.28
UEM–LOOKING	EMP	7.03	7.51	–0.48	0.29
	UEM – Layoff	1.03	0.65	0.38	0.26
	UEM – Looking	78.00	74.61	3.39*	0.21
	NLF	13.94	17.23	–3.29*	0.18

* Significant at $\alpha = 0.001$.

Table 2 shows the same set of estimates for the truly employed population only in somewhat greater detail. In this table, we considered the two primary subclassifications of unemployed: UEM-LAYOFF and UEM-LOOKING. This table provides information regarding the source difference in accuracy rates between the two questionnaire versions. We first consider the misclassification of true LAYOFF persons (top half of the table) and then consider the LOOKING persons (bottom half of the table).

For persons on layoff, classification accuracy appears to have dropped an average of 5.66 percentage points with the introduction of the revised questionnaire: from 61.30 percent to 55.63 percent. However, the patterns of classification error also changed. For the original questionnaire, the probability that a person on layoff is misclassified as looking for work is estimated at about 18 percent. The corresponding estimate for the revised questionnaire is less than half that: 8.5 percent. In addition, the data suggests that misclassification of unemployed persons on layoff as either employed or not in the labor force increased by 10.35 and 4.52 percentage points, respectively.

Now consider persons who are truly looking for work in the bottom half of Table 2. According to the MLCA model, classification accuracy for the redesigned CPS decreased significantly from 78.00 to 74.61 percent. Most of the misclassification is attributed to misclassifying persons

looking for work as NLF. This result would arise, for example, if the questions regarding active and passive job search activities are prone to error. To further investigate this finding, we conducted an analysis of each of the questions used to determined the LOOKING recode. In the next section, we first consider the sources of error in the LAYOFF classification and then investigate the sources of error for the LOOKING classification.

4.2 Specific Questions Responsible for the Reduction in LAYOFF Accuracy

4.2.1 Decomposition of the LAYOFF Recode

Individuals in the CPS are classified as LAYOFF on the basis of their responses to five questions in the original questionnaire and eight questions in the revised questionnaire. These questions are listed in Figure 3. Initially, we consider which questions or combinations of questions contribute most to the error rate observed in Table 2 for the LAYOFF recoded variable and then show how MLCA models can be applied to estimate the contributions to classification error of individual questions that are used to classify an individual as LAYOFF. The methodology employed for this is similar to the MLCA approach used previously for estimating the aggregate classification error. We will describe this technique in terms of the LAYOFF classification, but it will be applied subsequently to

decompose the error in both the LAYOFF and LOOKING classification processes.

First, we combine the questions in Figure 3 using the logical operators such as “and,” “or,” “if-then-else,” *etc.* to form a set of dichotomous “compound” questions with the property that each compound question must be answered positively in order for an individual to be classified as LAYOFF by the CPS classification process. Let $Q_k, k = 1, \dots, K$ denote the outcomes to the K compound questions that were formed for the LAYOFF classification, where $Q_k = 1$ denotes a positive outcome and $Q_k = 2$ denotes a negative outcome. Then an individual in the CPS is classified as LAYOFF if and only if $Q_k = 1$ for $k = 1, \dots, K$. In Figure 4, we define a set of four compound questions for original questionnaire, labeled O1–O4, and five compound questions for the revised questionnaire, labeled N1–N5.

For each classification, Q_k there is a corresponding true, unobservable (latent) classification, T_k defined in analogy to Q_k ; *i.e.*, an individual is truly on layoff by the CPS definition if and only if $T_k = 1, k = 1, \dots, K$. Next, we will use MLCA to estimate the misclassification error rates for each compound question Q_k by treating these as indicators for the unknown true latent characteristics, T_k .

The probability of an error in the classification of LAYOFF can be written as

$$\Pr(Q_k = 2 \text{ for some } k, k = 1, \dots, K \mid T_k = 1, k = 1, \dots, K) \quad (7)$$

which is the probability that an individual who is truly on layoff answers at least one the K compound questions negatively.

Next, we define the latent variable, W , as the number of compound questions for which the true response is positive, *i.e.*,

$$W = \begin{cases} 0 & \text{if } T_1 = 2, T_2 = 2, \dots, T_K = 2 \\ 1 & \text{if } T_1 = 1, T_2 = 2, \dots, T_K = 2 \\ \dots etc \dots & \\ K & \text{if } T_1 = 1, T_2 = 1, \dots, T_K = 1. \end{cases} \quad (8)$$

For example, $W = 0$ if a person’s true response pattern to the questions O1–O4 is (2,2,2,2), $W = 1$ if the true response pattern is (1,2,2,2), and so on. Note that $W = K$ corresponds to a true layoff. Thus, for the original questionnaire, $W = 0, \dots, 4$ and for the revised questionnaire, $W = 0, \dots, 5$.

To decomposing the probability in (7) into individual components for the compound question, Q_k , we rewrite (7) in terms of the error probabilities associated with each compound question. Thus, it can be shown that (7) can be rewritten as

$$\sum_{k=1}^K \Pr(Q_1 = 1, \dots, Q_{k-1} = 1, Q_k = 2 \mid W = K). \quad (9)$$

The k^{th} term in the sum may be interpreted as the contribution of question Q_k to probability of being misclassified given a true LAYOFF.

To estimate the components of (9) using MLCA, we define a classification variable, R , which is defined in analogy to W for the observed values of Q_k ; *i.e.*,

$$R = \begin{cases} 0 & \text{if } Q_1 = 2, Q_2 = 2, \dots, Q_K = 2 \\ 1 & \text{if } Q_1 = 1, Q_2 = 2, \dots, Q_K = 2 \\ \dots etc \dots & \\ K & \text{if } Q_1 = 1, Q_2 = 1, \dots, Q_K = 1. \end{cases} \quad (10)$$

Original Questionnaire	Question Wording
Q19	What were you doing most of LAST WEEK?
Q20	Did you do any work at all LAST WEEK not counting work around the house?
Q21	Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?
Q21A	Why were you absent from work LAST WEEK?
Q22E	Could you have taken a job LAST WEEK if one had been offered?
Revised Questionnaire	
Q20	LAST WEEK, did you do ANY work (either) for pay (or profit)?
Q20B-a	LAST WEEK, (in addition to the business,) did you have a job, either full or part time? Include any job from which you were temporarily absent.
Q20B-b	LAST WEEK, were you on layoff from a job?
Q20B-1	What was the main reason you were absent from work LAST WEEK?
Q21	Has you employer given you a date to return to work?
Q21A	Have you been given any indication that you will be recalled to work within the next 6 months?
Q21A-1	Could you have returned to work LAST WEEK if you had been recalled?
Q21A-2	Why is that?

Figure 3. Primary Components of UEM for the Original and Revised Questionnaires

Compound Question Number	Source Question(s) from the CPS Questionnaire	Compound Question Response is Positive if Source Question Response is....
Original Questionnaire		
O1	Q19: What were you doing most of LAST WEEK? or Q20: Did you do any work at all LAST WEEK not counting work around the house?	Q19: Any response except working and Q20: No
O2	Q21: Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?	Yes
O3	Q21A: Why were you absent from work LAST WEEK?	Temporary layoff (Under 30 days) or Indefinite layoff (30 days or more or no definite recall date)
O4	Q22E: Could you have taken a job LAST WEEK if one had been offered?	Yes
Revised Questionnaire		
N1	Q20: LAST WEEK, did you do ANY work (either) for pay (or profit)?	No
N2	Q20B-a: LAST WEEK, (in addition to the business,) did you have a job, either full or part time? Include any job from which you were temporarily absent.	Any response except "retired," "disabled", or "unable to work"
N3	Q20B-a: LAST WEEK, were you on layoff from a job? or Q20B-1: What was the main reason you were absent from work LAST WEEK?	Q20B-b: Yes Or Q20B-1: "On layoff" or "slack work/business conditions"
N4	Q21: Has your employer given you a date to return to work? or Q21A: Have you been given any indication that you will be recalled to work within the next 6 months?	Q21: Yes or Q21: No and Q21A: Yes
N5	Q21A-1: Could you have returned to work LAST WEEK if you had been recalled? or Q21A-2: Why is that?	Q21A-1: Yes or Q21A-1: No and Q21A-2: Own temporary illness

Figure 4. Compound Questions Used in the LAYOFF Recode for Original and Revised Questionnaire Versions

Let $\pi_{k|K}^{R|W}$ denote $\Pr(R = k | W = K)$. Then for $k > 0$ we may write

$$\pi_{k|K}^{R|W} = \Pr(Q_1 = 1, \dots, Q_{k-1} = 1, Q_k = 2 | W = K). \quad (11)$$

Thus, the contributions to error of each LAYOFF question can be obtained from the probabilities in (11).

To estimate the probabilities $\pi_{k|K}^{R|W}$ we fit MLCA models to the same data from the 1993 and 1994 CPS as used in the previous analysis and replicated the analysis on the 1993 parallel survey data. Data from the 1992 and 1995 CPS were not part of this analysis. The MLCA models used were similar to those described in the analysis for Tables 1 and 2. That is, we used three consecutive months of data and estimated the components in (10) for 10 consecutive, overlapping intervals for each year (*i.e.*, January–March, February–April, and so on to October–December). For the original questionnaire, the model specified three latent variables corresponding to the three months within a time period, each with $K + 1 = 5$ latent classes. For the revised questionnaire, we use an identical model except each latent variable had $K + 1 = 6$ latent classes.

As before, the best MLCA model for this analysis incorporated the proxy-self grouping variable, P , and specified non-stationary transitions, equal response

probabilities within time period, group heterogeneous transition probabilities, and heterogeneous response probabilities. The model provides an adequate fit to the data for all months in the analysis (*i.e.*, $p > 0.05$).

Table 3 provides a summary of the results from this analysis. In the column labeled "percent of total" we report $p_k \times 100$ percent where

$$p_k = \frac{\hat{\pi}_{k|K}^{R|W}}{\sum_{k=1}^K \hat{\pi}_{k|K}^{R|W}} \quad (12)$$

is the proportion of the classification error due to compound question k in Figure 4 and where $\hat{\pi}_{k|K}^{R|W}$ are the MLCA estimates of $\pi_{k|K}^{R|W}$.

The contribution to total error presented in Table 3 (Percent of Total column) is estimated by $p_k \times \Pr(A \neq 2 | X = 2)$ where p_k is given by (12) and $\Pr(A \neq 2 | X = 2)$ is estimated from Table 2 as 1 minus the accuracy rate for LAYOFF. For the original questionnaire, the components that contribute most to LAYOFF classification error are question O2 (64.2 percent) and question O1 (27.2 percent). These two questions taken together explain more than 90 percent of the error in the LAYOFF classification.

For the revised questionnaire, estimates from the 1994 CPS indicate that more than 90 percent of the error in the LAYOFF classification arises from two components: N1 and N4.

The analysis for the revised questionnaire was repeated on the Parallel Survey with very similar results. The same two components emerge as contributing more than 90 percent of the error. As mentioned in section 2, the utility of the 1993 Parallel Survey as an indicator of data quality for the revised questionnaire is in doubt. Nevertheless, the agreement of the results from the Parallel Survey and the 1994 CPS adds strength to the findings from the 1994 CPS analysis.

Thus, reduction in LAYOFF classification accuracy for the revised questionnaire appears to be due primarily to error in the responses to two compound questions: N1, the revised global question "LAST WEEK, did you do ANY work (either) for pay (or profit)?" and N4, which determines whether an individual reporting some type of layoff has a date or indication of a date to return to work. The MLCA estimates indicate that almost 60 percent of the error in the revised LAYOFF classification maybe attributed to N1 while about 34 percent may be attributed to N4.

4.2.2 Decomposition of the LOOKING Recode

The estimation process described for LAYOFF was also applied to the LOOKING recode. Note that compound question O1, O2, N1, and N2 defined in Figure 5 for LOOKING are the same questions as defined in Figure 4 for LAYOFF. Since O1, O2, and N1 appeared to be problematic for LAYOFF, we might expect that they might also be problematic for LOOKING.

Following the approach used for LAYOFF, for each survey year, we defined a latent variable, W in (8) and an indicator variable, R in (9). As we did in the LAYOFF analysis, we fit MLCA models to the data and determined that the best MLCA model for the analysis is the model

incorporating the proxy-self grouping variable, P , and specifying non-stationary transitions, equal response probabilities within time period, group heterogeneous transition probabilities, and heterogeneous response probabilities. This model provides an adequate fit to the data for all months in the analysis (*i.e.*, $p > 0.05$). As before, we include the results from the Parallel Survey for comparison with the 1994 CPS results; however, the latter results will be emphasized.

Table 4 displays the values of p_k defined in (11) for the LOOKING classification. For the original questionnaire, the major contributors to classification error appear to be questions O1 and O3, which contribute 31.5 and 56.3 percent of total classification error, respectively. Question O2, which was quite problematic for the LAYOFF population, appears less so for the LOOKING population. While it contributes 64.2 percent of the LAYOFF error estimate (or 24.8 percentage points to the error rate), O2 only contributes 11.3 percent of the LOOKING error estimate (or 2.5 percentage points to the error rate).

For the revised questionnaire, the results from the analysis of the Parallel Survey and the 1994 CPS are again quite similar. The component N1 appears to be an important source of error for LOOKING as it was for the LAYOFF analysis. However, its contribution to LOOKING is smaller: 10 percentage points compared with 25 percentage points for LAYOFF. The biggest contributor to LOOKING error seems to be question N3 which contributes 64.5 percent of the error based on the CCO analysis and 51.1 percent based on the 1994 CPS analysis.

Thus, the initial labor force question appears to be problematic for both questionnaire versions. The MLCA suggests that persons who are looking for work as well as persons who are on layoff experience some difficulty responding to the question "LAST WEEK, did you do ANY work (either) for pay (or profit)?" The changes made to this question in 1994 do not appear to have improved the accuracy of this question for the either population.

Table 3

Percent Contributions to Error in LAYOFF Classifications for Compound Questions for the 1993 CPS, Parallel Survey, and the 1994 CPS

Question	1993 CPS (Original Version)		Parallel Survey (Revised Version)		1994 CPS (Revised Version)	
	Error Rate	Percent of Total	Error Rate	Percent of Total	Error Rate	Percent of Total
Old Questionnaire						
O1	10.53	27.20	—	—	—	—
O2	24.84	64.19	—	—	—	—
O3	2.35	6.08	—	—	—	—
O4	0.67	1.74	—	—	—	—
New Questionnaire						
N1	—	—	23.19	52.26	25.34	57.12
N2	—	—	0.00	0.00	0.00	0.00
N3	—	—	2.76	6.22	3.06	6.90
N4	—	—	18.42	41.52	15.07	33.98
N5	—	—	0.00	0.00	0.89	2.00
Total	38.39	100.00	44.37	100.00	44.37	100.00

Table 4
Percent Contributions to Error in LOOKING Classifications by Compound Questions for the 1993 CPS, Parallel Survey, and the 1994 CPS

Question	1993 CPS (Original Version)		Parallel Survey (Revised Version)		1994 CPS (Revised Version)	
Old Questionnaire	Error Rate	Percent of Total	Error Rate	Percent of Total	Error Rate	Percent of Total
O1	6.93	31.51	—	—	—	—
O2	2.49	11.34	—	—	—	—
O3	12.39	56.33	—	—	—	—
O4	0.18	0.83	—	—	—	—
New Questionnaire						
N1	—	—	8.38	33.00	10.00	39.40
N2	—	—	0.00	0.00	0.00	0.00
N3	—	—	16.38	64.5	12.97	51.08
N4	—	—	0.46	1.81	2.27	8.96
N5	—	—	0.18	0.71	0.14	0.56
Total	22.00	100.00	25.39	100.00	25.39	100.00

Compound Question Number	Source Question(s) from the CPS Questionnaire	Compound Question Response is Positive if Source Question Response is....
Old Questionnaire		
O1	Q19: What were you doing most of LAST WEEK? or Q20: Did you do any work at all LAST WEEK not counting work around the house?	Q19: Any response except working and Q20: No
O2	Q21: Did you have a job or business from which you were temporarily absent or on layoff LAST WEEK?	No
O3	Q22: Has ... been looking for work during the past 4 weeks? and Q22A: What has ... been doing in the last 4 weeks to find work?	Q22: Yes or response to Q19 was LK (LOOKING) and Q22A: Response other than "nothing"
O4	Q22E: Could ... have taken a job LAST WEEK if one had been offered?	Yes or No, and reason is "Already has job" or "Own temporary illness"
New Questionnaire		
N1	Q20: LAST WEEK, did you do ANY work (either) for pay (opr profit)?	Q20: No
N2	Q20B-a: LAST WEEK, (in addition to the business,) did you have a job, either full or part time? Include any job from which you were temporarily absent.	Q20B-a: No ¹
N3	Q22: Have you been doing anything to find work during the last 4 weeks?	Yes
N4	Q22A: What are all the things you have done to find work during the last 4 weeks? Or Q22A-DK: You said you have been trying to find work. How did you go about looking? And Q22A-DK1: Can you tell me more about what you did to search for work?	Mention of at least 1 active activity.
N5	LAST WEEK, could you have started a job if one had been offered?	Yes

¹ Note: In a few cases, N2 was positive if response to Q20B-a was "Disabled" or "Unable" and response to Q20A-1: "Does your disability prevent you from accepting any kind of work during the next six months?" was "No".

Figure 5. Compound Questions Used in the LOOKING Recode for Original and Revised Questionnaire Versions

The key difficulty for the LOOKING category appears to be determining whether persons who are truly looking for work have made efforts of any type (either passive or active) in the past four weeks to find work. If a respondent is classified correctly as having made some effort, the next step in the process – *viz.*, determining whether the efforts satisfy the definition of active looking – is not problematic according to the estimates in Table 4.

5. CONCLUSIONS

Biemer and Bushery (2000) provides some evidence that unemployment classification accuracy rates in the 1994 CPS redesign survey were smaller than for the original survey design used prior to 1994. This paper provides additional evidence of their findings based upon a more extensive analysis of CPS data from 1992 through 1994. Our results

indicate that the probability of correctly classifying unemployed persons decreased from 79.1 percent to 73.5 percent – a difference of 5.6 percentage points. We estimate that roughly 60 percent of the reduction (3.4 percentage points) is due to an increase in the classification error for persons on layoff while the remainder (2.2 percentage points) is due to an increase in the classification error for persons looking for work.

For the revised questionnaire, both LAYOFF and the LOOKING classifications are each based upon five compound questions. For LAYOFF, two compound questions emerged as being problematic. One is the initial labor force question, which asks “LAST WEEK, did you do ANY work (either) for pay (or profit)?” The contribution of this component to LAYOFF misclassification is estimated to be approximately 57 percent which is more than double the corresponding rate for this question in the original questionnaire. In addition, a large error rate is estimated for the compound question formed by two questions: “Has your employer given you a date to return to work?” and “Have you been given any indication that you will be recalled to work within the next 6 months?” Approximately 34 percent of the estimated LAYOFF error rate is due to this combination. Since there are no corresponding questions in the original questionnaire, most of the error in classifying persons on layoff in the revised questionnaire may be linked to these two questions.

For classifying persons who are looking for work in the redesigned survey, two questionnaire components appear to contribute most to classification error: “LAST WEEK, did you do ANY work (either) for pay (or profit)?” and “Have you been doing anything to find work during the last 4 weeks?/What has...been doing in the last 4 weeks to find work?” The error rates for both questions are slightly larger for the revised questionnaire than for the original questionnaire. These increases, therefore, explain the slight increase in LOOKING classification error observed for the revised questionnaire.

The error in CPS unemployment classification is well-documented; for example, see Chua and Fuller 1987; Abowd and Zellner 1985; Porterba and Summers 1995; and Sinclair and Gastwirth 1998. A widely accepted measure of reliability for the CPS – viz., index of inconsistency computed CPS reinterview – shows the reliability of the CPS unemployment classification decreased after the redesign. Results provided in this paper are consistent with these prior studies and help determine the source of the error in the CPS classification of the unemployed. At a minimum, our results provide a basis for further investigation into the root causes of the errors in the collection of labor force data in the CPS. Through cognitive laboratory experiments and field experiments, we may identify causes of the error in the

unemployment questions that would suggest ways to improve the questions. Such improvements could be implemented in a future redesign of the CPS.

ACKNOWLEDGEMENT

The author would like to acknowledge the assistance of Pamela McGovern at the U.S. Census Bureau who commented on early drafts of the paper. Appreciation is also expressed to the Associate Editor and an anonymous referee, both of whom were very helpful in preparing the article. Financial support for this research was provided by the U.S. Census Bureau.

REFERENCES

- ABOWD, J. and ZELLNER, A. (1985). Estimating gross Labor-Force flows. *Journal of Business and Economic Statistics*, 3, 3, 254-283.
- BIEMER, P. (2004). Modeling measurement error to identify flawed questions. In *Methods for Testing and Evaluating Survey Questionnaires*, (Eds. S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin and E. Singer), Hoboken, New Jersey: John Wiley & Sons, Inc., 225-246.
- BIEMER, P., and BUSHERY, J. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, 26, 139-152.
- BIEMER, P., and WIESEN, C. (2002). Latent class analysis of embedded repeated measurements: An application to the National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society, Series A*, 165, 1.
- CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 397, 46-51.
- COHANY, S., POLIVKA, A. and ROTHGEB, J. (1994). Revisions Current Population Survey. Employment and Earnings BLS Report.
- COHEN, J.A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 37-46, 1960.
- DIPPO, C., POLIVKA, A., CREIGHTON, K., KOSTANICH, D. and ROTHGEB, J. (1994). Redesigning a Questionnaire for Computer-Assisted Data Collection: The Current Population Survey Experience.
- KOSTANICH, D., and CAHOON, L. (1994). CPS Bridge Team Technical Report 3: Effect of Design Differences Between the Parallel Survey and the New CPS. Unpublished Report.
- MILLER, S. (1994). What Would the Unemployment Rate Have Been Had the Redesigned Current Population Survey Been in Place from September 1992 to December 1993?: A Measurement Error Analysis. CPS Bridge Team Technical Report 1.
- POLIVKA, A. (1994). Comparisons of Labor Force Estimates from the Parallel Survey and the CPS During 1993: Major Labor Force Estimates. CPS Overlap Analysis Team Technical Report 1.

- POTERBA, J., and SUMMERS, L. (1995). Unemployment benefits and labor market transitions: A multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.
- POULSEN, C.S. (1982). *Latent Structure Analysis with Choice Modeling Applications*. Doctoral dissertation, Wharton School, University of Pennsylvania.
- ROTHGEB, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality, U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, April 6.
- SINCLAIR, M., and GASTWIRTH, J. (1998). Estimates of the Errors in classification in the Labour Force Survey and their effects on the reported unemployment rate. *Survey Methodology*, 24, 157-169.
- THOMPSON, J. (1994). Mode Effects Analysis of Labor Force Estimates. CPS Overlap Analysis Team Technical Report 3.
- VAN DE POL, F., and DE LEEUW, J. (1986). A Latent Markov Model to Correct for Measurement Error. *Sociological Methods and Research*, 15, 1-2, 118-141.
- VERMUNT, J. (1997). *ℓ EM: A General Program for the Analysis of Categorical Data*, Tilburg, University.
- WIGGINS, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*, Elsevier S.P.C., Amsterdam.

Comment

JEROEN K. VERMUNT¹

1. INTRODUCTION

I enjoyed very much reading this very well written paper. The topic addressed by Paul Biemer – classification errors in the measurement of employment status – is a very important one. Employment statistics belong to the most important macro-economic indicators and, actually, we would wish they would be free of error. It, however, turns out to be impossible to measure a person's employment without error. The best that can be done is design the data collection in such a manner that the classification errors at the individual level are minimized as much as possible. The current paper contributes to this objective.

An earlier study by Biemer and Bushery (2000) indicated that the 1993 changes in the measurement procedure that intended to reduce classification errors actually increased measurement error. In the current paper, Paul Biemer replicates these former analyses with a longer time series and with an extra employment category obtained by splitting the unemployed group into "on layoff" and "looking for work". The reported results confirm the earlier conclusions that the new procedure is worse than the old procedure. In a second step, Biemer tries to disentangle the sources of measurement error for the two unemployed categories by modeling the separate questions that are used to determine whether a person is "on layoff" and "looking for work", respectively. Sources of error are identified that point at possible improvements in the questionnaire.

Because of my background, my commentary will mainly concern methodological and statistical issues. More precisely, I will discuss some methodological problem related to application of the LC Markov model, as well as indicate how the statistical analysis could be somewhat refined. It is, however, not clear whether such a more elegant modeling will yield very different conclusions. I want to stress ones more that this is a great paper. My critical remarks are only meant to stimulate the discussion.

2. LATENT CLASS MARKOV: METHODOLOGY

The main engine of the study performed by Paul Biemer is the LC of hidden Markov model. Several assumptions that may affect the encountered results have to be made

when – as in this study – the model is applied with a single indicator per occasion. The assumption that is discussed in detail by Biemer is the first-order Markov process assumption. Simulation studies by Biemer and Bushery showed that, fortunately, estimates of classification error are not very sensitive to this assumption. Another assumption that is needed here for model identification is that the measurement error is constant over time. This assumption does not seem to be very problematic in the current study since we are looking for a single time-constant measure for classification error. Moreover, there is no good reason to assume that the quality of the measurement procedure changed over time while the procedure itself did not change (of course, apart from the questionnaire redesign). I am much more concerned about the third assumption; that is, the assumption of independent classification errors (ICE) over time (Bassi, Hagenaars, Croon and Vermunt 2000). Is it realistic to assume that the occurrence of a certain type of classification error at time point t does not affect the probability of making the same mistake at time point $t + 1$? In my opinion, this assumption is not realistic in the current application. For example, a respondent who makes a mistake because (s)he did not understand one of the questions will most probably (or at least be more likely than others) make the same error again at the next occasion. In my opinion, it is necessary to conduct a simulation study to determine the sensitivity of the estimated classification errors for violations of the ICE assumption.

I have another critical remark concerning the use of the LC Markov model for quantifying measurement error in a person's employment state. According to the model, there is a probabilistic relationship between an individual's true and observed states. What is, however, the true state? Is it the true employment state occupied at a particular time point, or the state that would have been recorded with an error-free or gold-standard instrument? Or is it the state a person would have occupied under "normal conditions"? That is, if also randomness in his/her behavior is filtered out.

I will illustrate my point with a small example. Suppose that there is two types (two latent segments) of coffee consumers: consumers who prefer brand A and consumers who prefer brand B, and that I belong to the brand B segment, which means that under normal circumstances I buy brand B coffee. In an interview, I am asked which

¹ Jeroen K. Vermunt, Department of Methodology and Statistics, Tilburg University, The Netherlands.

brand I bought last week. Suppose I report that I bought a brand A package of coffee, and that am neither lying nor making a mistake. In other words, there is no classification error in the sense of making a mistake: I really bought brand A this week (the researcher doesn't know that of course). On the other hand, my behavior from this week is inconsistent with my preference, which means that in terms of measurement of my preference there is a classification error. This example illustrates that there are two types of "errors" that can be made: an error in the reporting and an "error" in the behavior. The "error" in my behavior of this week may have many causes, such as "brand B was sold out", "brand A was offered at a lower price this week", "I could not find the brand B package because of changes in the arrangement of the supermarket", *etc.* The LC Markov model is not able to distinguish such randomness in the behavior that is uncorrelated across time points from real classification errors.

What does this imply for the employment application? It implies that an individual's true state may be "on layoff", but for some reason (by chance) this particular month (s)he has worked. If this "some reason" is uncorrelated with other "some reasons" for being in the "wrong" observed state at other occasions, it will be labeled classification error by the LC Markov model. While in the case of the measurement of preferences based on revealed (or stated) preferences correcting for randomness in behavior seems to be exactly what we wish to accomplish, this is clearly not the case in the measurement of employment status. I, therefore, have the strong feeling that the error rates reported by Biemer might be somewhat overestimated because of randomness in employment behavior, for instance, caused by randomness in the functioning of the labor market.

A well-known consequence of modeling individual change by means of a LC Markov model is that the estimated number of latent transitions is much smaller than the corresponding observed numbers. The reason for this is that both independent classification errors and independent random behavior is filtered out; that is, part of the observed change is attributed to these phenomena.

3. LATENT CLASS MARKOV: MODEL SPECIFICATION

Paul Biemer estimated a separate three-occasion LC Markov model for each of the 30 three-month data sets. Interview mode was used as a grouping variable in order to take into account some of the heterogeneity in the true employment distributions and classification errors. The reported error rates in the tables are averages over interview modes and rotation groups.

I would have set up the model in a somewhat more elegant and less ad hoc manner. Instead of running a separate analysis for each of the rotation groups, I would have tried to build a simultaneous model for all rotation groups. The main problem of doing a series of separate analyses is that parameters that should actually be equated across rotation groups are now estimated without constraints. For example, the employment distribution in March 1994 should be the same in the rotation groups that were interviewed between January and March, February and April, and March and May, respectively. Moreover, the transition probabilities between March and April should be the same in the February–April and March–May rotation groups. This has also implications for the Parallel Survey groups: their time-specific latent distributions and transitions should be assumed to be equal to the ones of the standard CPS. That would have been a much better manner to test whether measurement error differ between the two questionnaires. Especially for the period in which the questionnaire forms overlap, it is crucial to assume equal latent distributions in order to be able to prevent that differences in measurement error appear partially as differences in true states.

A similar problem of the separate analyses applies for the estimation of the classification errors. These are assumed to be time-constant within the 3-month period that a rotation group is interviewed, but are allowed to differ across rotation groups, even if they are interviewed in the same month. It would, of course, be much better to impose equality constraints across rotation groups. A consistent application of the time-homogeneity assumption would imply that – both for the old and the new questionnaire form – the measurement errors are constant within the full investigation period.

What we, actually, need is a LC Markov model covering all 30 months; that is, a model for 30 instead of 3 time points. Such a simultaneous model for all rotation groups is as easily specified as a model for 3 time points. Of course, for each rotation group, only 3 of the 30 months are observed, which means that the other time points have to be treated as missing values. This is not a problem in the maximum likelihood estimation of the model parameters since we can simply assume that the data are missing at random (Vermunt 1997). Questionnaire type (old/new) serves as grouping variable (in addition to interview mode) and affects the time-homogenous classification error probabilities. In other words, we estimate only two sets of classifications errors, one for the old and one for the new questionnaire. Transition probabilities may change over time, but will be equal across rotation groups interviewed at the same occasions. Moreover, the initial state probabilities of a rotation group are not estimated as separate parameters

since they are defined by the current state of the latent Markov chain.

A practical problem of the simultaneous modeling is that with so many time points it no longer possible to estimate the model parameters with the standard EM algorithm. With a variant of EM called the Baum-Welch algorithm, however, the model can also be applied with many time points (Vermunt 2003; Paas, Bijmolt and Vermunt 2003). This algorithm is implemented in an experimental version of the Latent GOLD program (Vermunt and Magidson 2000, 2003) and will be available in a next version of this program.

An alternative way to implement a simultaneous model is as a LC Markov model for 3 occasions in which rotation group serves as grouping variable and in which the relevant across rotation group equality restrictions are imposed on the classification errors, transition probabilities, and initial state probabilities. The most complicated part of this approach is that it requires the use of restrictions on marginal probabilities (Vermunt, Rodrigo and Ato-Garcia 2001). More precisely, the initial state probabilities should be in agreement with the marginal class sizes in the rotation groups that are interviewed at the same occasion.

Other aspects of the modeling that could be refined are the treatment of missing values and the coding of the interview mode. It is not necessary to eliminate cases with missing values from the analysis as is done by Paul Biemer because ML estimation with missing values is straightforward. As far as the interview mode is concerned, it would be much more elegant to work with only two categories – proxy and self – instead of four categories and let the interview mode vary across occasions within cases. In other words, interview mode could be used as a time-varying covariate. Vermunt, Langeheine and Böckenholt (1999) proposed such a latent class Markov model with time-varying covariates.

4. MODEL FOR RESPONSE PROCESS

It is a very nice idea to try to disentangle which questions in the questionnaire are causing the classification errors by modeling the response process itself. This may yield lots of valuable information for redesigning the questionnaire. I, however, think that the extended models for the employment statuses “on layoff” and “looking for work” are formulated in an overly complicated manner.

The form of the created variable R is the same as of the outcome variable in a sequential choice analysis or in a discrete-time survival analysis. Answering the next question is fully determined by whether the current one is answered positively or not. The information we have is how many steps a person takes, which is conceptually equivalent to a

discrete survival time. A person “surviving” till the end is classified as being “on layoff” (“looking for work”).

In my opinion, it is not very helpful to treat this variable as being generated by K latent variables (Ts). This only makes sense if theoretically there should be a response hierarchy at the latent level, which, however, because of measurement error, is not encountered at the manifest level. That is, if at the manifest level there are 2^K instead of K possible responses. Even if is the case, it often suffices to conceptualize the model as a model with a latent variable with $K + 1$ classes and K indicators, a structure that is sometimes referred to as a probabilistic Guttman model.

Paul Biemer recognizes the complexity of the K latent and K manifest variables formulation and decides to simplify the model. However, I assume because of his starting point, he decided to keep $K + 1$ latent classes. I do not see why so many latent classes are needed. There are not even so many employment states. More logical would be to have only two classes – “on layoff” and “not on layoff” (“looking for work” and “not looking for work”) – since the questions are only intended to make this particular distinction. It can, of course, happen that the questions turn out to be informative about the type of “not on layoff” (“not looking for work”) status, in which case an extra latent class might be needed. What is clear to me is that $K + 1$ classes are far too many.

I was wondering how many persons were classified as “on layoff” (“looking for work”) at the various time points in the analysis with composite variable R as indicator. Are these numbers, as well as the number of transitions into and out of this state similar to the ones obtained with the standard four-state LC Markov model. In my opinion, this is a requisite for the validity of the calculation performed to obtain the figures presented in Tables 3 and 4.

A final thing that occurred to me is the following. Why not building a LC Markov model using the full questionnaire information as is done in the second part of the analysis. In other words, an alternative to using the observed constructed classification consisting of 4 employment categories would be to use the full set of CPS employment questions answered by the respondents. Such an analysis with multiple indicators would not only be much more informative, it would also make it possible to test and relax some of the assumptions that were made in the current analysis. For example, the ICE assumption could be relaxed for some of the questionnaire items.

REFERENCES

- BASSI, F., HAGENAARS, J.A., CROON, M. and VERMUNT, J.K. (2000). Estimating true changes when categorical panel data are affected by uncorrelated and correlated classification errors. *Sociological Methods and Research*, 29, 230-268.

- PAAS, L.J., BIJMOLT, T.H. and VERMUNT, J.K. (2003). Extending dynamic Segmentation with Lead Generation: A Latent Class Markov Approach. Center Paper, Tilburg University (submitted for publication).
- VERMUNT, J.K. (1997). *Log-linear models for event histories*. Techniques in the Social Sciences Series, Thousand Oakes: Sage Publications. 8.
- VERMUNT, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33. In press.
- VERMUNT, J.K., LANGEHEINE, R. and BÖCKENHOLT, U. (1999). Latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24, 178-205.
- VERMUNT, J.K., and MAGIDSON, J. (2000). *Latent GOLD User's Manual*. Boston: Statistical Innovations Inc.
- VERMUNT, J.K., and MAGIDSON, J. (2003). *Addendum to Latent GOLD User's Guide: Upgrade for Version 3.0*. Boston: Statistical Innovations Inc.
- VERMUNT, J.K., RODRIGO, M.F. and ATO-GARCIA, M. (2001) Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods and Research*, 30, 170-196.

Comment

STEPHEN M. MILLER and ANNE E. POLIVKA¹

1. INTRODUCTION

We are grateful for the opportunity to comment on this interesting paper. We will focus most of our comments on the empirical findings about the 1994 Current Population Survey (CPS) redesign, rather than a technical discussion of the Markov Latent Class Analysis (MLCA) methodology itself.

In his article, "An Analysis of Classification Error for the Revised Current Population Survey Employment Questions," the author applies MLCA models in an effort to trace the source of what he believes to be the "reduced accuracy of the revised classification of unemployed persons" after the redesign. In the CPS individuals are considered to be unemployed either because they are classified as being on layoff or because they are classified as looking for work. The author reports a particularly large reduction in the accuracy of the measurement of persons on layoff. Consequently, we will focus our attention on the classification of individuals on layoff, although similar comments can be made about the change in the measurement of those looking for work. In examining the accuracy of the measurement of those on layoff, the author assumes that those classified as on layoff were conceptually the same before and after the 1994 redesign, and that these individuals should exhibit the same labor force flows month-to-month. There are, however, many reasons why the improved measurement embodied in the redesign should conceptually change who is classified as on layoff. In addition, there are several factors unrelated to changes in question wording that could affect the composition of those classified as on layoff. Therefore, what the author describes as a reduction in accuracy due to the redesign more appropriately could be attributed to conceptual changes in those classified as on layoff, and the fact that what was being measured by the CPS before the redesign is not the same as what is being measured by the CPS after the redesign.

2. IMPROVED MEASUREMENT

One of the main reasons for the CPS redesign was to more accurately measure official definitions and concepts.

Layoff was found to be an especially problematic concept, in that its meaning in general usage in the 1990's – a permanent job separation – was very different from the official CPS definition – a temporary job separation with the expectation of recall. When the questions were originally written in the 1940's, the term layoff was commonly used to refer to temporary spells of unemployment due to retooling or slowing of business conditions. Consequently, recall expectations were not asked about in the pre-redesign questionnaire. Research conducted in the 1980s and early 1990s in preparation for the redesign indicated that respondents' interpretation of layoff had become considerably broader than the official definition. Focus group interviews and large scale respondent debriefings found that between 30 and 50 percent of those who said they were on layoff did not expect to return to their former employers (Rothgeb 1982; Palmisano 1989; Polivka and Rothgeb 1993). Also, in 1993, 5.4 percent of those classified as on layoff had last worked 1 to 5 years ago, and another 0.6 percent had not worked in the last 5 years. This lack of recent work experience further supports the notion that many of those classified as on layoff prior to the redesign had no expectation of recall.

To better measure the official CPS definition of layoff, two questions were added in the revised questionnaire asking about individuals' recall expectations – "Has your employer given you a date to return to work?" and "Have you been given any indication that you will be recalled to work within the next 6 months?" Individuals for whom the answer is "yes" to either of these questions are classified as on layoff if they are available for work; all others are excluded from being classified as on layoff (these individuals can be classified as unemployed later in the questionnaire if they meet the active job search and availability criteria).

As a result of the addition of these direct questions, a somewhat different group of people would be expected to be classified as on layoff. Prior to the redesign, a substantial proportion, if not the majority, of individuals classified as on layoff were in fact permanently separated from their employers. After the redesign, those classified as on layoff had to expect to be recalled to their former employers; thus the vast majority of these individuals should be only temporarily separated from their employers. It is not

¹ Stephen M. Miller and Anne E. Polivka, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Room 4985, Washington, D.C. 20012.

surprising that these two groups of individuals would exhibit different month-to-month flows between labor force groups. It is reasonable to expect that individuals who expect to be recalled to their job would be more likely than those who are permanently separated to go from being temporarily on layoff to employed in consecutive months. Further, compared with permanently separated workers, those in industries in which temporary layoffs are prevalent would be more likely to be on layoff one month, employed the next month, and then laid off again.

Month-to-month gross flows of individuals between labor force states indicate that there was an increase in the proportion of the unemployed who went to employment after the 1994 redesign. Specifically, in 1994, 26.6 percent of those who were unemployed in the first month were employed in the second month, compared with 23.7 percent in 1993.

The author's MLCA estimates of a supposed decrease in the accuracy of those classified as on layoff after the redesign because more individuals are classified as employed subsequent to being on layoff, in reality is exactly in accord with what would be expected with a tightening of the definition of on layoff, and is consistent with the increase in the month-to-month gross flows between unemployment and employment (although the increased flow also is in accord with a declining unemployment rate that was observed during the time period covered by the author's study). The MLCA's smaller, but still significant, estimated decrease in accuracy due to more individuals on layoff being classified as not in the labor force after the redesign also is consistent with the tightening of the definition of on layoff through the requirement that individuals expect to be recalled in the next six months, given that individuals may adapt or change their recall expectations over time. For instance, when first interviewed, individuals may expect to be recalled in the next six months. However, in subsequent months, as the time from the initial separation increases, these individuals may no longer say that they expect to be recalled. If, at the same time, these individuals have not started searching for alternative employment, perhaps because they are still eligible to receive unemployment insurance payments, these individuals would transition to being not in the labor force. Alternatively, individuals may initially expect to be recalled; however, in subsequent months due either to poor weather conditions or a deteriorating economic situation for their former employers these individuals may become more uncertain about the probability of being recalled and thus they may not say that they expect to be recalled. If in later months, economic conditions for their former employers improve or the weather becomes less inclement, these individuals again may correctly feel that they will be recalled. The existence of

changing expectations could generate a three month pattern where individuals truly were on layoff in the first month, not in the labor force the second month, and on layoff again in the third month. Those who were permanently separated from a job and were incorrectly classified as on layoff in the unredesigned survey would be unaffected by changing recall expectations. Consequently, individuals who were permanently separated from their jobs probably would be more likely to report themselves as on layoff in consecutive months with the unredesigned survey. The MLCA model would interpret this greater stability as indicating that those on layoff were more accurately measured prior to the redesign. However this greater "accuracy" would only be amongst those who were incorrectly classified because they used too broad a definition.

The author concludes that 60 percent of the misclassification of those on layoff in the redesigned survey is due to the question "LAST WEEK, did you do ANY work for pay?" This actually is consistent with more people being on temporary layoff and being recalled by their former employers in the redesigned survey (although if individuals on layoff engage in temporary employment while waiting to be recalled to their former employers, an increase in transitions to employment after 1994 may also be at least partially attributable to the broader employment question used in the redesigned survey). Similarly, the author concludes that 40 percent of the misclassification of those on layoff in the redesigned survey is due to the expectation of recall questions ("Has your employer given you a date to return to work?" and "Have you been given any indication that you will be recalled to work within the next 6 months?"). This is consistent with changing recall expectations and a slight increase in the flow between on layoff and not in the labor force. The author is obtaining different MLCA estimates of those classified as on layoff before and after the redesign because the composition of those groups has been changed, and the composition of the groups have changed in a manner that was desired and intended by those who redesigned the questionnaire.

Further evidence of the different composition of those classified as on layoff can be found in a comparison of data that were collected to determine the effect of the redesign on labor force estimates generated from the CPS. Prior to January 1994, the redesigned questionnaire was administered to 12,000 households monthly from late 1992 to December 1993. After the new questionnaire was implemented in 1994, the old questionnaire was administered monthly from January 1994 to May 1994 to 12,000 households drawn from the same sample. The experimental administration of the old and redesigned questionnaires has been referred to as the "Parallel Survey". Parallel Survey estimates from before 1994 using the new methodology and

after 1994 using the old methodology were generated to compare to official CPS estimates using the unredesigned CPS procedures prior to 1994 and the redesigned procedures after 1994. Polivka and Miller (1998) illustrate the importance of using both parts of the Parallel Survey to obtain a complete picture of the effects of the redesigned survey. For instance, if just the first part of the Parallel Survey were used, it would have been estimated that the redesign increased the unemployment rate by 0.5 percentage point. In fact, when both parts of the Parallel Survey were used, the redesign was estimated to have no statistically significant effect on the unemployment rate.

Using both parts of the Parallel Survey and the official CPS estimates, Polivka and Miller estimate that the redesigned CPS decreased the proportion of unemployed men who were on layoff by a little less than 7 percent, while it increased the proportion of unemployed women classified as on layoff by almost 7 percent (although the latter estimate was not statistically significant at a 5 percent level). These estimates imply that the redesign would decrease the proportion of those on layoff who were male and increase the proportion who were female compared to the proportions that were obtained prior to the redesign, if all else were equal. Comparison of annual averages for those over the age of 20 support this notion, since they indicate that, in 1993, 67.2 percent of those on layoff were male, compared to 63.6 percent of those on layoff in 1994 (although in addition to questionnaire changes these proportions could be affected by changes in economic conditions).

The industry distribution of those classified as on layoff, using data from both parts of the Parallel Survey and the official CPS, reveals other compositional changes in those classified as on layoff before and after the redesign. Examination of estimates from the redesigned survey to the official CPS estimates for January to May 1993 and from the unredesigned survey to official CPS estimates for January to May 1994 reveals particularly dramatic differences for those in the durable manufacturing industry. The proportion of those on layoff who were formerly employed in durable manufacturing when the unredesigned questions were used was almost half the proportion obtained when the redesigned questions were used (for January to May 1993 the proportion of those on layoff who were formerly employed in durable manufacturing averaged 16.8 percent among those who received the unredesigned questions and 9.8 percent among those who received the redesigned questions. For January to May 1994 the proportions were 8.7 percent among those who received the unredesigned questions and 15.5 percent for those who received the redesigned questions). At the same time the proportion of those on layoff who were in construction was 10 to 15 percent larger when the redesigned questions were used

compared to when the unredesigned questions were used (for January to May 1993 the proportion of those on layoff who were formerly employed in the construction industry averaged 33.3 percent for those who received the redesigned questions and 27.4 percent for those who received the unredesigned questions. For January to May 1994 the proportions were 33.3 percent and 25.9 percent respectively).

Averaging the average difference between the first part of the Parallel Survey and the CPS for January 1993 to May 1993 (which is equal to the new method effect plus the Parallel Survey effect) with the average difference between the CPS and the second part of the Parallel Survey for January 1994 to May 1994 (which is equal to the new method effect minus the Parallel Survey effect) indicates that the redesign decreased the proportion of those classified as on layoff who were formerly employed in the durable manufacturing industry by 7.3 percentage points and increased the proportion classified as formerly employed in the construction industry by 3.7 percentage points (averaging the average difference between the first part of the Parallel Survey and the CPS with the average difference between the CPS and the second part of the parallel survey is in the spirit, albeit a simplified version, of the main-effects linear models estimates using generalized least squares that were presented in Polivka and Miller).

Individuals in different industries could have very different true labor force transition patterns which in turn could be influencing the MLCA estimates. For instance, given that a substantial proportion of employment in the construction industry is sensitive to weather conditions and may be more project-oriented than other types of employment, it is not unreasonable to expect that workers in construction might truly be more likely to be temporarily laid off in the first of three consecutive months, employed on a short term basis in the second month (either because the weather improved in the second month or because a short term construction project was undertaken), and then temporarily laid off again in the third month (either because weather conditions deteriorated or the project for which they were hired was completed). On the other hand, employment in the durable manufacturing industry has been steadily declining since the 1970's (for example, comparing non-recession years, it was estimated that in 1971 14.9 percent of U.S. workers as measured by BLS's establishment survey were employed in the durable manufacturing industry, compared to 9.2 percent in 1993 and 8.5 percent in 2000). This long term decline in employment makes it likely that a large proportion of workers in the manufacturing industry classified as "on layoff" prior to the redesign were permanently separated from their employers (the change in the industry distribution when the expectation of being recalled was imposed is consistent with this notion). Being

permanently separated from a job in combination with the relatively high wages workers in durable manufacturing received may increase the likelihood of these individuals being unemployed in three consecutive months, because it takes time to find employment in another industry at a similar wage.

Comparison of MLCA model estimates before and after the redesign without accounting for differences in industry composition of those classified as on layoff could cause analysts to mistakenly conclude that the redesign decreased the accuracy of labor force classifications. In reality, the increase in transitions that were measured after the redesign represented a true increase in transitions to employment after layoff was properly asked about in the CPS questionnaire. Failure to account for the fact that the redesigned CPS questionnaire intentionally classified a somewhat different group of individuals on layoff than did the unredesigned questionnaire could lead to incorrect conclusions being drawn from the MLCA models. Workers permanently separated from their employers who were classified as on layoff using the unredesigned questions are appearing to be more accurately classified in MLCA models, but they are more stable in a classification that was incorrect in the first place. Further, a proportion of individuals who are correctly classified as on layoff according to the official definition inherently could have less stable employment histories due either to their personal tastes or the industries with which they are associated.

In addition to compositional changes related to differences in question wording, the author also may have inadvertently captured in his estimates several other compositional changes unrelated to wording differences. These include differences in the time periods the author used for his estimates, as well as technological changes in the data collection process and economic conditions.

3. SEASONALITY

The first inadvertent compositional difference the author may have introduced is related to seasonality and the different time frames the author used for estimation. The number of individuals classified as on layoff in the CPS has a great deal of seasonal variability, with typically a larger number of individuals being on layoff early in the year. For instance, there were 358 individuals who were classified as on layoff in January 1995 who matched to February and March, while there were 294 individuals classified as on layoff in March 1995 who matched to April and May, and only 188 people classified as on layoff in June 1995 who matched to July and August. This means that there were 18 percent more people initially classified as on layoff in January 1995 than in March 1995 and 47 percent more

individuals classified as initially on layoff in January 1995 than in June 1995. Using three month moving averages generated with the same calendar months probably would help to mitigate the effects of seasonality. However, the author did not use the same monthly time spans to generate his three-month moving averages to estimate the MLCA models before and after the redesign. The majority of the author's pre-redesign estimates were generated using data from August 1992 through December 1993, while the majority of his post-redesign estimates were generated using data from January 1994 to May 1995. Using these time spans means that the author only has, for instance, one January to March matched set of data for the pre-redesign estimates, while he has two January to March matched sets of data for the post-redesign estimates.

4. TECHNOLOGICAL CHANGES IN DATA COLLECTION

A second reason that the composition of the groups in various labor force states may be different for data collected with the unredesigned and the redesigned methodology is related to the ability to match individuals' data from month to month and the quality of these matches. The vast majority of data collected using the unredesigned methodology either in the official CPS prior to January 1993 or in the Parallel Survey from January 1994 to May 1994 were recorded using a paper form, and interviewers were required to transcribe by hand household and person identification numbers from master files to the paper survey forms. All of the data collected using the redesign methodology, either in the official CPS after January 1994 or in the Parallel Survey in 1993, were collected using an automated instrument that was loaded onto either a laptop computer or on a centralized computer. As part of the computerized data collection process, household and person identification numbers were automatically and consistently carried forward month to month. Using paper forms and transcribing data by hand has the potential to introduce errors and cause researchers to eliminate as non-matches individuals who actually are the same individuals and thus true matches.

Using the same public-use data that the author used, in combination with additional information about whether an individual had moved (that is periodically collected in the CPS), Madrian and Lefgren (1999) estimated that, depending on the stringency of the match criterion used, between 64 and 87 percent of those who were eliminated as an invalid match probably legitimately did match. Further, Madrian and Lefgren noted that there was a substantial decline between 1993 and 1996 in the fraction of invalid matches that probably should have been retained in the data set based on the criterion of whether an individual had

moved (since Madrian and Lefgren were using publicly released data, they were not able to investigate the validity of matches for 1994 to 1995 and 1995 to 1996 because the ability to match this data was suppressed to protect individuals' confidentiality). Madrian and Lefgren suggest that the increased number of valid matches for 1996 onward was due to improvements attributable to the redesign (it should be noted that, although a better match can be obtained using data internal to BLS and the Census Bureau in which information has not been suppressed, the quality of a match using internal data still will be affected by the data collection methodology. Thus the quality of the match will be better after the redesign than before the redesign). In their research, Madrian and Lefgren also found that individuals who were incorrectly excluded from the matched data sets were much more likely to be young and have their information provided by another member of the household (a proxy responder). These individuals are also the ones that Biemer argues are more likely to have classification errors in their labor force status. Consequently, by potentially including more of these individuals in his study due to the improved quality of the match, the author could be obtaining a decrease in the accuracy of his measures that he incorrectly is attributing to the questionnaire.

5. ECONOMIC CONDITIONS

Economic conditions may also contribute to differences in the composition of the groups classified as on layoff before and after the redesign. From 1992 to 1995, the period which the author uses for the majority of his MLCA modeling, the unemployment rate was steadily declining. Specifically, in 1992 the annual average unemployment rate was 7.5 percent while in 1995 it was 5.6 percent.

At a higher unemployment rate, it is likely that the proportion of individuals who remain unemployed month to month is larger than at lower unemployment rates. As the economy improves and the unemployment rate declines, it is not unreasonable to expect an increase in the proportion of individuals who transition from being on layoff to employment. With the increase in these transitions to employment, the proportion of individuals who transition to temporary jobs might also increase. Indeed, although undoubtedly related to many factors, the number of individuals employed in the temporary help supply industry (as defined under the NAICS coding system) increased 44 percent between 1992 and 1995 – from 1.1 percent to 1.5 percent of the U.S. establishments' payrolls (as measured by the BLS's establishment survey).

In addition, as the unemployment rate declines, the type of individual classified as unemployed may change.

Specifically, those who remain unemployed when the unemployment rate is low tend to find it more difficult to become steadily employed and are more likely to transition quickly between labor force states. This is the logic behind studies that analyze the effects of different types of employment separations on subsequent labor force outcomes. For instance, in a study comparing individuals who were separated from their employers due to slack business conditions as opposed to complete plant shut downs, Gibbons and Katz (1991) found that, with regard to both duration of joblessness and earnings, workers who were separated from their employers due to slack business conditions did significantly worse than did those who were separated due to a plant closing. Gibbons and Katz argue that these differences were due to employers being able to dismiss their least productive workers, while retaining their more productive workers, when business conditions were slack, as opposed to employers having to dismiss both their least productive and most productive workers when a plant was completely shut down. Similarly, Darby, Haltiwanger and Plant (1985) argue that as economic conditions worsen, the duration of unemployment increases as a result of a change in the composition of those who are unemployed. This is because in more adverse economic conditions, the proportion of the unemployed who are high-skill workers (who also are less used to being unemployed and more likely to be able and willing to hold out for a more satisfactory job) will increase and the proportion of the unemployed who are less skilled and who frequently transition between labor force states will decrease.

It is important to note that the majority of the author's pre-redesign estimates were generated using 1992 and 1993 data, when the unemployment rate averaged 7.0 percent, while the majority of the redesigned estimates were generated using data from 1994 and 1995, when the unemployment rate averaged 6.0 percent. Changes in general economic conditions, and corresponding changes in the composition of the unemployed, may be affecting the supposed accuracy of the author's estimates in a way that is unrelated to the questionnaire. For instance, between 1992 and 1995, the proportion of the unemployed who were teenagers steadily increased from 14.8 percent to 18.2 percent, while the overall unemployment rate steadily declined from 7.5 percent to 5.6 percent. Similarly, the proportion of the unemployed who were Hispanic steadily increased from 13.6 percent to 15.4 percent between 1992 and 1995, though some of this may be due to the increasing proportion of Hispanics in the population (which rose from 8.8 percent to 9.4 percent). Both teenagers and Hispanics tend to be lower skilled workers who historically have been more likely to become unemployed or withdraw from the labor market. It should be noted that, regardless of the

source, an increase in the proportion of the unemployed drawn from groups with less stable labor force histories will influence the MLCA model estimates of accuracy if the change is not accounted for in the modeling.

6. DIFFERENTIAL VALIDITY OF THE MARKOV ASSUMPTIONS

In addition to differences in the composition of those classified as on layoff affecting the estimates generated by the MLCA models, differences in the composition of the various labor force groups before and after the redesign could affect the validity of the underlying assumptions of the MLCA models. As the author notes, a key assumption when implementing MLCA models is that an individual's transition from the second to third month is independent and thus uninfluenced by how the individual was classified in the first month. When estimating MLCA models for individuals' labor force states this obviously is untrue, and the validity of the assumption will likely differ amongst the various labor force categories. For instance, an individual who is employed in the first month is much more likely to be employed in the third month than is an individual who has never worked. More importantly, an individual cannot be classified as on layoff in either the redesigned or unredesigned questionnaire if he or she has not previously worked. Addition, under the official definition of layoff that was implemented in the redesign, individuals also have to expect to be recalled. This leads to a much tighter relationship between employers and workers across months using the redesigned questionnaire. Given that individuals on layoff under the redesign are much more likely to be recalled and thus employed than under the unredesigned questionnaire, the likelihood of an individual's labor force status in the third month depending on their initial labor force status in the first month is much higher. Consequently, not only is it likely that the Markov assumptions are often violated in labor force studies; it is much more likely that the Markov assumptions are violated after the redesign. This differential violation of the model's assumptions could be fundamentally influencing the author's results.

7. CONCLUSION

In summary, although the author believes that he identified a problem that was introduced into the CPS with the 1994 redesign, the supposed increase in misclassification of those on layoff in reality reflects the greater

precision of the survey questions. Rather than identify a true error, we believe the author may have failed to recognize that the composition of the groups identified as on layoff before and after the redesign were different due to both intentional changes (such as the definition of on layoff being built into the questionnaire or improved quality of matches obtained because of computerization of the survey) and to uncontrolled changes such as developments in the overall economy. Finally, we would like to see further work in this area which combines the MLCA modeling approach along with a careful consideration of the economic concepts being measured, the time periods being examined and the assumptions being made. We believe this could lead to a more accurate understanding of the effects of the 1994 CPS redesign, and more useful application of the MLCA modeling approach in general.

ACKNOWLEDGEMENTS

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics. The authors would like to thank Sharon Cohany, U.S. Bureau of Labor Statistics, for helpful commentary on this discussion.

REFERENCES

- DARBY, M.R., HALTIWANGER, J. and PLANT, M. (1985). Unemployment rate dynamics and persistent unemployment under rational expectations. *American Economic Review*, 75, 614-637.
- GIBBONS, R., and KATZ, L.F. (1991). Layoffs and Lemons. *Journal of Labor Economics*, 9, 351-380.
- MADRIAN, B.C., and LEFGREN, L.J. (1999). A Note on Longitudinally Matching Current Population Survey (CPS) Respondents. Technical Working Paper 247, *National Bureau of Economic Research Technical Working Paper Series*.
- PALMISANO, M. (1989). Respondents' Understanding of Key Labor Force Concepts Used in the CPS. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA.
- POLIVKA, A.E., and MILLER, S.M. (1998). The CPS After the Redesign: Refocusing the Economic Lens. In *Labor Statistics Measurement Issues*, (Eds., J. Haltiwanger, M.E. Manser and R. Topel), National Bureau of Economic Research Studies in Income and Wealth, Chicago: University of Chicago Press, 60, 249-286.
- POLIVKA, A.E., and ROTHGEB, J. (1993). Overhauling the Current Population Survey: Redesigning the Questionnaire. *Monthly Labor Review*, 116, 10-28.
- ROTHGEB, J. (1982). Summary Report of July Follow-up of the Unemployed. U.S. Bureau of the Census Memorandum, Washington D.C.

Comment

CLYDE TUCKER¹

1. INTRODUCTION

I first would like to congratulate Paul Biemer for offering an innovative approach to the study of measurement error in surveys. Although he chose to illustrate his approach with the employment series in the Current Population Survey (CPS), the method can be applied to many surveys. My comments largely will be conceptual in nature, but I will supplement these comments with examples from the same data that Biemer analyzed.

Using Markov Latent Class Analysis (MLCA), the Biemer paper relies on an evaluation of the consistency over time of respondents' answers to the questions in the employment series. The increase in inconsistency found in the new series as compared to the old one, after controlling for self versus proxy reports, may serve as an indicator of one type of measurement error in the assignment of labor force status. Presumably, this error is the result of the failure of the new questions (at least, compared to the old ones) to collect the correct information for classifying an individual into the right labor force category. Thus, the error can be attributed to poor question design. Because the analysis indicates that the errors tend to be in one direction more than in the other – the misclassification of truly unemployed individuals into a different category – some might interpret the result to be a bias in the unemployment rate.

I will argue that not only has bias not been introduced but also that the new series, while certainly not perfect, reduces error, providing a more accurate picture of the employment situation. It does this by taking into account the economic realities of today in a way that the old series did not. This is accomplished by not only better question wording but also by the inclusion of follow-up questions and probes that capture more detailed information for determining a respondent's true employment status. The use of follow-up questions and probes is facilitated by the introduction of a computerized survey instrument. As a result of these innovations, I believe that the new employment series reduces the amount of specification error that existed with the old series. By specification error, I mean the error arising from using questions that do not measure what they are intended to measure. I also will explain why I do not believe that Biemer's method is appropriate for use in this particular case.

2. RECOGNITION OF THE NEED FOR A NEW EMPLOYMENT SERIES

The last major revision of the CPS prior to 1994 took place in 1967. In the ensuing years, the labor market underwent a great transformation. The number of women in the labor force dramatically increased. The number of part-time jobs and multiple job holdings escalated. The relationship between the worker and the employer became more tenuous. Startling technological developments changed the way Americans did work and resulted in the creation of new types of jobs requiring new kinds of skills. Perhaps most importantly, the economy gradually became more service oriented and less manufacturing oriented.

Just one result of these developments that needed to be taken into account in the CPS was the change in the accepted meaning of "layoff" as so ably described by Miller and Polivka (2004), but there were others, as enumerated by Bregger and Diplo (1993). Better information was needed about discouraged workers (those who have given up looking for work), multiple jobholders, marginal workers (e.g., unpaid workers in a family business), and job-changing patterns. In addition, during the 1970s and 1980s, concern mounted about the various types of nonsampling errors that could be affecting CPS estimates as well as about respondent burden and its detrimental effect on data quality.

Until the 1980s, the technology to tackle these problems was not available. However, as Bregger and Diplo (pages 4–5) note, things began to change:

"...in the early 1980s, the introduction of two new survey methodologies provided the means for understanding and reducing measurement error. These included the application of behavioral science methods and theory – more commonly referred to as the cognitive aspects of survey methodology – and computer-assisted interviewing. It is through the blending of these two methodologies that a new collection procedure, which focuses on reducing measurement error, was made possible."

Cognitive methods (including focus groups and in-depth interviewing) made it possible to develop questions that

¹ Clyde Tucker, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 1950, Washington, D.C. 20212.

could accurately measure the more complex economic behaviors that the times required. Furthermore, these techniques were able to uncover problems in the existing labor force series (See Polivka and Rothgeb 1993). The accurate measurement of the more complex behaviors also required a more complicated survey instrument. One so complicated that interviewers, left to their own devices, would have difficulty navigating. This is where computer-assisted interviewing played an important role. With a computerized survey instrument, interviewers could easily navigate through the complex skip patterns necessary to obtain answers to questions for measuring the wide variety of economic behaviors of interest.

3. CONSIDERATION OF NONSAMPLING ERRORS IN BOTH THE OLD AND NEW CPS EMPLOYMENT SERIES

Let me begin this section by detailing my reasons why MLCA is not an effective tool for evaluating the new CPS design relative to the old one. MLCA can be a good method for detecting measurement error within a constant series of questions by looking for inconsistencies in response over several administrations to the same respondent. In the case of the CPS, the method might be appropriate, given a careful examination of a well-chosen set of diagnostics, for examining problems in the old employment series and the new employment series independently of one another. However, let me add a caveat here about examining inconsistencies even within the same employment series. Labor force status, in itself, is inherently inconsistent over time. While the employed and not-in-the-labor-force (NILF) categories are relatively stable, the unemployed category is not. Those in that category are trying to get out. Controlling for seasonal effects by looking at March-May of either 1993 or 1994, it turns out that, on average, almost 90% of those in the employed and NILF categories did not move from one month to the next. On the other hand, over half of those in the unemployed category did. Thus, the unemployed are a particularly difficult group for MLCA to handle.

As for comparing the two series, the use of MLCA is problematic because the two series were designed to measure different things. There were some significant changes made in the employment series in the hopes of reducing specification error. Although I do not want to dwell on the measurement of layoff (Miller and Polivka have covered this topic well.), I do want to use it as a case in point for explaining why the comparison of the old and new instrument is a difficult one to make. Apart from what Miller and Polivka have said, I have my own reasons for doubting Biemer's conclusions.

The changes in the layoff questions were designed to reduce the specification error discovered in qualitative research on the meaning of "layoff," as alluded to by Miller and Polivka. In the attempt to eliminate specification error, two additional questions were added. One asked whether a date for recall had been given, and the other inquired about the possibility of returning to the job within the next 6 months. Only those who were given a recall date or expected to return to work within the 6-month period were classified as truly "on layoff."

Clearly, this altered the characteristics of the group classified as unemployed as a result of layoff as well as those asked the remaining questions in the employment series, but I believe there also were more subtle reasons why inconsistencies in respondents' answers could have increased and still not have contributed to measurement error to the extent argued by Biemer. In the first place, respondents had to answer more questions, which would have increased the probability that at least one false inconsistency would arise from one month to another. This might add to measurement error compared to the old series, but specification error, considered to be the greater problem, still would be reduced. Furthermore, false inconsistencies arising from these questions should be minimized for two reasons. These questions are much more specific than the single layoff question in the old series, and they had been well tested (Esposito, Campanelli, Rothgeb and Polivka 1991). Moreover, given that more specific questions were asked, there would be an increased chance that true change had taken place in the state of at least one of them in the intervening month. Finally, and of greatest interest to me, is the fact that these questions attempt to capture information on relatively nuanced changes. For instance, a respondent may have changed his or her mind about the possibility of being recalled in the next 6 months based on little concrete information. With the uncertainties in today's job market, it would be difficult to say that the respondent had given the wrong answer.

I now want to address Biemer's concerns about the initial question in the new employment series asking about whether any work was done last week for "either pay or profit." His results indicate that this question may be contributing to the amount of error he finds in both the "layoff" and "looking" series. The change in this question (as well as the addition of a question on the existence of a family-owned business or farm) was prompted by the concern that the old questions were not stated broadly enough, so that marginal workers, especially those working for profit at home, were not being classified as working. For example, the Parallel Survey showed the percentage of part-time workers in the new CPS was 1.098 times larger than in the old CPS, and, coincidentally, the employment to

population ratio for women 65 and older also increased by about the same amount (Polivka and Miller 1998). The same is true when comparing 1993 to 1994. It stands to reason that the increased precision in the identification of these marginal workers, who are more likely to be inconsistent in their answers from month to month than other workers, might be mistaken for measurement error. The fact is the more narrow "what were you doing last week" question could lead these respondents to consistently, but inaccurately, report they were unemployed.

Finally, let me turn to the other section of the employment series in which Biemer found a problem – the "looking for work" questions. One important change in this series involved clarifying the differences in "active" and "passive" job search in order to reduce misclassification rates in these categories. Studies conducted in the 1980s found that interviewers were confused about what constituted an active (versus a passive) job search (Polivka and Rothgeb 1993). In the redesigned questionnaire, interviewers were given an explicit list of both active and passive job search methods.

Comparisons of the results of the old and new questions are complicated by the fact that different subpopulations were asked these questions in the two series. Those finally defined as looking (and, thus, considered unemployed) in the two different employment series could have arrived there in quite different ways. Half of those considered looking in 1993 received that designation by volunteering they were looking in the first question ("What were you doing most of last week?"); none of those who were looking in 1994 followed that path. Those retired and 50 or older in 1994 never got the chance to say they were looking. In 1993, none of those who said they were on layoff were asked the looking question, so they had no chance to be classified as NILF in a given month. Then there were the two different levels of information given to the interviewers for coding active and passive methods. One difference uncovered in an analysis of the two groups from 1993 and 1994 was that a higher proportion of those looking in 1994 were women compared to 1993 (45.4% vs. 41.2%). Referring to the above discussion on the first employment question, increases in the inconsistency in reports to the looking questions could be the result of capturing more marginal workers using the revised employment series. Sometimes these individuals would be looking and sometimes not.

4. CONCLUSIONS

Paul Biemer has made a bold attempt to investigate the error structure in the CPS employment series; however, his findings do not take into account the reasons for the revised questions. Taking these into account would help explain the month-to-month inconsistencies that he found. Not only might these inconsistencies be real, but they could provide evidence of a reduction in specification error. For instance, controls other than for self/proxy could be included in the model to take into account some of the changes in methodology, and measurement error within more limited subpopulations. More exploration of the utility of MLCA with inherently inconsistent classifications also should be undertaken.

ACKNOWLEDGEMENTS

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics. The author would like to thank Steve Miller, Anne Polivka, and John Dixon for their assistance on this discussion.

REFERENCES

- BREGGER, J.E., and DIPPO, C.S. (1993). Overhauling the Current Population Survey: Why is it necessary to change? *Monthly Labor Review*, 116, 3-9.
- ESPOSITO, J.L., CAMPANELLI, P.C., ROTHGEB, J.M. and POLIVKA, A.E. (1991). Determining which questions are best: Methodologies for evaluating survey questions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 46-55.
- MILLER, S.M., and POLIVKA, A.E. (2004). Discussion of the paper An analysis of classification error for the revised Current Population Survey employment questions. *Survey Methodology*, 30, 145-150.
- POLIVKA, A.E., and MILLER, S. (1998). The CPS after the redesign: Refocusing the economic lens. In *Labor Statistics Measurement Issues*, (Eds. J. Haltiwanger, M.E. Manser, and R. Topel). National Bureau of Economic Research Studies in Income and Wealth. Chicago: University of Chicago Press, 60, 249-286.
- POLIVKA, A.E., and ROTHGEB, J. (1993). Overhauling the Current Population Survey: Redesigning the questionnaire. *Monthly Labor Review*, 116, 10-28.

Response from the Author

PAUL P. BIEMER¹

1. INTRODUCTION

My sincere thanks to all four discussants for their thoughtful, thorough and constructive comments. They have added considerably to our understanding of the complex issues surrounding Markov Latent Class Analysis (MLCA) and the Current Population Survey (CPS) labor statistics. All four discussants raise a number of important issues that I will try to address to the extent I can. Some issues will require more work and deserve much greater consideration than is possible here. More complete responses to those issues will have to await the results of future research.

Considering all the comments collectively, there seems to be agreement that Markov latent class analysis has considerable potential as a tool for evaluating and exploring the sources of measurement error in the CPS. However, there is some skepticism that it has identified real problems in the CPS questionnaire. Dr. Vermunt, who is also the author of the software I used for this analysis (*viz.*, *ℓEM*), provides a number of valuable suggestions for improving the models and investigating the validity of the model assumptions. The three other reviewers (Drs. Miller, Polivka, and Tucker) are quite familiar with the CPS since they are employed by the federal agency that sponsors the survey where they played important roles in the 1994 redesign. Their comments remonstrate the various ways in which the MLCA model assumptions could be violated for these data. In addition, they contain valuable information regarding details of the CPS (both pre- and post-redesign) and the construction of the CPS labor force variable. The comments and suggestions of all the discussants should be carefully considered by labor force economists and statisticians who are conducting research in the area of employment measurement error, particularly those using MLCA.

JEROEN VERMUNT'S COMMENTS

I first address the comments of Dr. Vermunt and then the comments of the other three reviewers. I share Dr. Vermunt's concern that the ICE assumption may not hold for these data. As he points out, if respondents

misunderstand the labor force questions in the same way from one month to the next, they may make the same errors each month creating correlated errors across the months. As an example, a person who is truly in the UEM category at both Times 1 and 2 may be more likely to be misclassified at Time 2 if they were also misclassified at Time 1. This can be stated probabilistically as

$$\rho = \frac{P(B \neq 2 | A \neq 2 \text{ and } X = Y = 2)}{P(B \neq 2 | A = 2 \text{ and } X = Y = 2)} - 1 > 0. \quad (1)$$

The numerator probability of the quantity ρ is the probability that the Time 2 classification (B) is in error given the Time 1 classification (A) is also in error and the true classification at both time points is UEM. The denominator probability is similar except for the condition that no error is made at Time 1 (*i.e.*, $A = 2$). Under the ICE assumption, $\rho = 0$. Therefore, if the $\rho > 0$ (which is the likely direction of the correlated error), the ICE assumption is violated. Dr. Vermunt suggests a simulation study be conducted to study the sensitivity of the estimated classification errors to violations of this assumption. Of course, determining the extent to which the ICE assumption fails for the CPS data is not possible via simulation. Nevertheless, it is still useful for assessing the potential for correlated error to bias the MLCA classification error estimates.

Following his suggestion, I conducted a small simulation study to gain some insight as to the consequences $\rho > 0$ for MLCA using CPS data. A sequence of artificial populations was generated using parameters consistent with those for the CPS (see for example, Table 1 in the main paper) except that ρ was increased in small increments from 0 to its empirical maximum – *i.e.*, the largest value of ρ that is feasible without violating the other model assumptions. Maintaining the other model assumptions in the analysis is necessary so that the consequences of violating just the ICE assumption can be isolated.

The largest feasible value of ρ was determined empirically to be 0.7. At this value of ρ , the MLCA estimate of the probability a correct classification of UEM went from 79% to 85% and the misclassification error rate dropped from 21% to 15%. For mild departures from the ICE assumption,

¹ Paul P. Biemer 3040 Cornwallis Road, PO Box 12194 Research Triangle Park, NC 27709-2194, U.S.A.

say $0 < \rho < 0.3$, the error rates changed by less than 3 percentage points. These results illustrate that if the ICE assumption fails to hold due to positive between interview correlations, the error rates estimated by MLCA will be somewhat underestimated. However, mild departures from the ICE assumption should have little effect on the classification error probabilities for these data. A similar analysis was conducted for the two other labor force categories (*i.e.*, EMP and NLF) but the change in the classification error estimates was negligible. This result was anticipated due to the relatively small error rates for these categories.

The results suggest that mild departures from the ICE assumption should have little or no effect the conclusions of the analysis. Extreme departures might affect the conclusions in the unlikely event that errors are highly correlated for original questionnaire and essentially uncorrelated for the revised questionnaire. Under that scenario, the original questionnaire would appear to have smaller UEM classification error than the revised questionnaire. However, there is no practical reason to expect this condition to hold since both questionnaires present questions that respondents may misunderstand consistently across interviews.

Although these simulation results, as well as those in Biemer and Bushery (2001) for investigating the consequences of violations of the Markov assumption, are quite useful for studying the sensitivity of the estimates to violations of the MLCA model assumptions, they provide no direct evidence of the validity of the MLCA estimates. Biemer and Bushery (2001) illustrate how the (empirical) validity of latent class estimates can be established using external data and alternative approaches for estimating classification error. A similar analysis based upon test-retest reinterview data will be provided in the sequel.

For the purpose of identifying potential areas where the CPS questionnaire can be improved, it is not essential to establish unequivocally that the MLCA model assumptions hold since model validity is of secondary importance. Instead, the primary issue for questionnaire evaluation work is whether the method of analysis used is successful at identifying questions that have large measurement errors and are in need of revision. In other words, the validity of the model is established by its ability to find important flaws in the questionnaire. Determining whether there truly is error in the UEM classification as suggested by MLCA requires an evaluation using other methods such as cognitive laboratory research. Cognitive interviews could be used to investigate encoding, comprehension, recall, and/or social desirability issues that generate errors in the responses to the UEM questions. If these investigations uncover important problems in questions, then the utility of MLCA for identifying flawed questions will be supported even

though the validity of the MLCA modeling assumptions may never be known.

Dr. Vermunt's other suggestions on ways the modeling framework could be improved are quite reasonable and I hope to investigate them further in the future. However, the current software for fitting MLCA models is somewhat limited and the estimation of complex models such as those he suggests may not be feasible. He also notes that problems can arise when fitting large models with the EM algorithm. As an example, initially we attempted to use the proxy/self-response variable as a time-varying covariate in the MLCA models, but encountered problems in the estimation process such as "division by 0" errors and persistent convergence to local maxima. We ultimately had to abandon the approach in favor of the single, time invariant proxy/self grouping variable used in the current analysis. As new and more general software becomes available, the options for MLCA with time varying covariates as well as other model enhancements mentioned by Dr. Vermunt will be feasible.

COMMENTS OF THE BLS DISCUSSANTS

I will address the comments of Drs. Miller and Polivka and those of Dr. Tucker together since the reviewers are from the same agency (BLS) and their comments raise similar concerns about the analysis. The following five points summarized their main concerns:

1. The modifications introduced in the new questionnaire capture more transitions than the old questionnaire. MLCA wrongly interprets these as errors when in fact they are not error.
2. Respondents may change their minds from month to month about whether their employers truly indicated that they might be recalled to work. These changes should not be classified as a response error.
3. The Markov assumption does not hold in labor force studies and it is violated to an even greater extent after the redesign than before the redesign. This differential violation of the model's assumptions could be fundamentally influencing the MLCA results.
4. The differences in the estimates of LAYOFF classification error before and after the redesign are due to the composition of the groups comprising this category. This composition changed after the redesign in a manner that was desired and intended by those who redesigned the questionnaire.

5. The increased inconsistency in reports to the LOOKING questions for the revised questions could be explained by more marginal workers being identified using the revised questions. Sometimes these individuals would truly be looking for work and sometimes not. MLCA misinterprets these ostensibly random changes as response error when they are not.

Point 1 describes an issue that should not pose any difficulties for MLCA. The MLCA model assumes that each individual occupies a true labor force state which may change from month to month. No assumption is made that the transition probabilities are the same for both questionnaires. The true initial labor force probabilities as well as the month-to-month transition probabilities are estimated independently for each questionnaire. In fact, although not discussed in main paper, the model estimates of the true exit probabilities for LOOKING and LAYOFF are in fact greater for the revised questionnaire than for the original questionnaire. Thus, a greater number of flows from one labor category to another for the revised questionnaire does not necessarily bias the estimates of classification error for that category in either direction.

Point 2 suggests that whether an individual is truly on layoff depends upon that individual's opinion about whether he or she was given an indication of possibly being recalled. However, this is not how the revised questionnaire defines the concept. An individual's true layoff status depends upon whether or not the employer truly provided an indication of being recalled. Although the respondent's opinion about what the employer indicated may change from month to month, the true layoff status does not change according to the respondent's opinion. Flows in and out of the LAYOFF category due to the respondent's opinion should be interpreted as error by the model.

Points 3, 4, and 5 could be made for any analysis employing MLCA. They essentially concern the potential bias in the MLCA estimates when month-to-month transitions do not behave according to the MLCA model and consequently real changes are misinterpreted as classification errors. As the reviewers note, there are at least three ways this can occur:

- a) the Markov assumption does not hold (point 3),
- b) there is unobserved or unexplained heterogeneity in the population (point 4), and
- c) employment-related behaviors for two consecutive months are not correlated for some persons; thus, for those persons, past month status does not predict the current month's status (point 5 as well as a point made by Dr. Vermunt).

The implications of (a) were considered in a simulation analysis in Biemer and Bushery (2001). Their results suggest that, for the CPS data, the estimates of classification error are quite robust to violations of the Markov assumption. It is unlikely, then, that non-Markov transitions explain the findings of higher classification error for the revised questionnaire. Still, additional research is needed to more thoroughly understand the implications of non-Markov transitions for our results.

For (b), it is quite possible for MLCA estimates to be biased when the compositions of the unemployed populations are substantially different under the original and revised questionnaires and those differences are not explained by the grouping variables used in the model. Likewise (c) may be regarded as a special case of (b). For (c), the transition probabilities for some population subgroup are uncorrelated with the prior month's employment status; instead it is correlated with other *unobserved* variables. In Jeroen Vermunt's coffee drinker example, the unobserved variable is the availability of a specific brand of coffee at the market. At this stage of the research, we have not conducted simulation studies to quantify the effects of unobserved heterogeneity on the estimates, but this possibility will be examined in future work.

However, this issue as well as the general plausibility of the MLCA estimates can be investigated to some extent by comparing the MLCA estimates with independent estimates from an estimation approach that is not affected by (a) through (c). If the findings from the alternative analysis are consistent with the MLCA findings, the MLCA findings gain credibility. As an example, test-retest reliability for the CPS employment classifications can be estimated both pre- and post-redesign using the CPS reinterview data (see for example Biemer and Forsman 1992 for a description of CPS reinterview program and these data). The validity of the estimates of test-retest reliability does not depend upon the Markov assumption or group homogeneity assumption; the ICE assumption, however, is still relevant for reliability estimation.

Table 1 shows estimates of Cohen's kappa measure of reliability for three time periods: 1992–1993, 1995–1997, and 2002–2003. As shown in the table, the reliability of the CPS classifications of unemployment dropped after the redesign from about 68% to 65%. The most recent estimates of kappa indicate reliability has dropped to below 60%. These results are consistent with the results from the MLCA that classification error in the CPS unemployment statistics has worsened after the redesign. It is possible that the reliability estimates in Table 1 are biased since they also rely on the validity of the ICE assumption. But as discussed previously, in order to the results in the table to be explained by the failure of the ICE assumption, the ICE assumption

would have to hold for the revised questions but not for the original questions. That condition is very unlikely to occur.

Table 1

Estimates of Cohen's Kappa for the CPS Before and After the Redesign

Year	<i>n</i>	Cohen's κ
1992 – 1993 ¹	28,063	67.8
1995 – 1997 ²	22,429	64.6
2002 – 2003 ³	19,205	58.8

¹ From Biemer and Bushery 2000.

² Bushery and McGovern (1999).

³ Personal communication with Bac Tran at the U.S. Census Bureau

Given the evidence presented here and in the main paper, it seems reasonable to consider the possibility that CPS unemployment classification error increased after the redesign. The next step is to conduct additional research to evaluate these findings and explore the possible causes for the error. Rather than to focus on the validity of the MLCA or test-retest reinterview models, the focus of the future research should be the revised CPS questions, particularly those used in the LAYOFF classification.

I have already mentioned the possibility of using cognitive interviews to investigating the problems in the response process associated with the revised questions. As an example, one question identified in the MLCA as being potentially flawed is: "Have you been given any indication that you will be recalled to work within the next 6 months?" Some of the issues that could be investigated in the cognitive laboratory for this question include:

- How well do unemployed subjects understand the meanings of terms such as "any indication" and "recalled?"
- Do subjects who were recently separated from employment have difficulty remembering what their employers said about being recalled when they were terminated?
- An employer may say, "If business improves, we may call you." Do respondents answer the question correctly in this situation?
- Do respondents who initially respond that they will be recalled later change their responses to this question as the months pass by and they have not been recalled?

SPECIFICATION ERROR AND MEASUREMENT ERROR

Finally, I will address an important issue raised by Dr. Tucker regarding specification error, measurement error and their net effects. As Dr. Tucker explains, the original questionnaire suffered from specification error bias caused

by measuring the wrong concept. The revisions to the labor force questions introduced in 1994 were designed to eliminate the specification error bias by refining the concepts of employment and unemployment and modifying the survey questions to reflect these refinements. These modifications, while reducing specification error, added more complexity to the survey questions which could have increased the measurement error bias in the labor force estimates. Dr. Tucker suggests that while this may be the case, the measurement bias in the new employment series may be less than the combination of specification bias and measurement bias in the old series. To determine whether this could be true, the specification error bias (B_S) and measurement error bias (B_M) were separately estimated using the MLCA estimates provided in the paper as described below.

Let p denote the CPS estimate of UEM and let P denote the expectation of p with respect to sampling and measurement error distributions. Let π denote the true value of the characteristic under the definitions of UEM implied by the specific questionnaire (*i.e.*, without regard to possible specification error). Therefore, $\pi = P - B_M$, *i.e.*, the value of P in the absence of measurement error bias.

As noted above, specification error bias is the bias in P due to a wrong concept or definition of unemployment implied by the questions and/or labor force classification process. For the revised questionnaire design, we assume that the specification error in p is 0 since it will be regarded as the gold standard for estimating the specification error bias in the original questionnaire.

Let π_{old} and π_{new} denote the π -parameter for the original and revised questionnaires, respectively. Then the specification error bias in the pre-1994 estimates of the unemployment rate is

$$B_S = \pi_{\text{old}} - \pi_{\text{new}}. \quad (2)$$

For each questionnaire, the estimate of P is p , the weighted estimate from the CPS. The estimate of π is obtained by correcting p for classification error bias using the response probabilities from the MLCA. Let $\mathbf{p}' = (p_1, p_2, p_3)$ where p_1, p_2, p_3 denote the estimates of the proportions in EMP, UEM, and NLF, respectively. Let ω_{ij} be the probability that an observation that truly belongs to the i^{th} category is assigned to the j^{th} category and let π_i denote the true proportion in the population in the i^{th} category. Then

$$E(\mathbf{p}) = \Omega' \boldsymbol{\pi} \quad (3)$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$ and $\Omega = [\omega_{ij}]$ is the 3×3 matrix with elements ω_{ij} . It follows that an estimator of $\boldsymbol{\pi}$ is

$$\hat{\boldsymbol{\pi}} = (\hat{\Omega})^{-1} \mathbf{p} \quad (4)$$

where $\hat{\Omega}$ is a MLCA estimate of Ω . For each questionnaire, $\hat{\Omega}$ was estimated by the average of the 10 MLCA estimates (January–March through October–December) using the 1993 CPS for the original questionnaire and 1993 Parallel Survey for the revised questionnaire.

Table 2 shows the results of this analysis. For UEM, $p = 6.38$ for the original and 6.98 for the revised questionnaire. If the unemployment rates are corrected for measurement bias using (4), unemployment rate increases to 7.09 percent for the original questionnaire and 8.03 percent for the revised questionnaire. Thus, an estimate of the measurement bias for the original survey is $6.38 - 7.09 = -0.71$ and for the revised survey is $6.98 - 8.03 = -1.05$. Note that the measurement biases are negative for both the original and revised questionnaires, indicating that UEM as well is underestimated by both questionnaire versions.

For the revised questionnaire, the specification bias is assumed to be 0. For the original questionnaire, it is estimated by the difference $7.09 - 8.03 = -0.94$ percent. An estimate of the net bias, $B_T = B_M + B_S$, is $-0.71 + (-0.94) = -1.65$ percent for the old series compared with $-1.05 + 0 = -1.05$ percent for the new series. Thus, while it is subject to greater measurement error bias, the new series has smaller estimated net bias assuming $B_S = 0$.

Several limitations of these results should be mentioned. First, as noted in the main paper, the estimates for revised questionnaire from the Parallel Survey may not be representative of the revised CPS series. Second, the

analysis assumes that the revised questionnaire is the gold standard for estimating the specification error bias in the original questionnaire. This assumption could also be challenged. Finally, no standard errors were provided for the estimates in Table 2 and the hypothesis of smaller overall bias in the revised question was not formally tested. Despite these limitations, the results suggest the possibility that the new unemployment series could have substantially lower net bias than the old series.

Table 2

Comparison of Original and Revised Questionnaire Biases for the CPS Unemployment Rate Based Upon Estimates from the 1993 CPS and the Parallel Survey

	p	π	B_M	B_S	B_T
1993 CPS	6.38	7.09	-0.71	-0.94	-1.65
Parallel Survey	6.98	8.03	-1.05	0 ¹	-1.05

¹Note: Specification error bias is assumed to be 0 for the revised questions.

REFERENCES

- BIEMER, P., and BUSHERY, J. (2001). Application of markov latent class analysis to the CPS. *Survey Methodology*, 26, 2, 136-152.
- BIEMER, P.P., and FORSMAN, G. (1992). On the quality of reinterview data with applications to the current population survey. *Journal of the American Statistical Association*, 87, 420, 915-923.

A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations

PATRICIA GUNNING and JANE M. HORGAN¹

ABSTRACT

A simple and practicable algorithm for constructing stratum boundaries in such a way that the coefficients of variation are equal in each stratum is derived for positively skewed populations. The new algorithm is shown to compare favourably with the cumulative root frequency method (Dalenius and Hodges 1957) and the Lavallée and Hidiroglou (1988) approximation method for estimating the optimum stratum boundaries.

KEY WORDS: Efficiency; Geometric progression; Neyman allocation; Stratification.

1. INTRODUCTION

A stratified random sampling design is a sampling plan in which a population is divided into mutually exclusive strata, and simple random samples are drawn from each stratum independently. The essential objective of stratification is to construct strata to allow for efficient estimation. In what follows X represents the known stratification or auxiliary variable while Y represents the unknown study variable. Suppose there are L strata, containing N_h elements from which a sample of size n_h is to be chosen independently from each stratum ($1 \leq h \leq L$). We write $N = \sum_{h=1}^L N_h$ and $n = \sum_{h=1}^L n_h$. In the case of the stratified mean estimate,

$$\bar{y}_{st} = \sum_{h=1}^L \frac{N_h}{N} \bar{y}_h, \quad (1)$$

where \bar{y}_h is the mean of the sample elements in the h^{th} stratum, we need to choose the breaks in order to minimise its variance

$$V(\bar{y}_{st}) = \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_{yh}^2}{n_h}, \quad (2)$$

where

$$S_{yh} = \sqrt{\sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 / N_h},$$

is the standard deviation of Y restricted to stratum and h , and

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi},$$

is the mean.

Dalenius (1950) derived equations for determining boundaries when stratifying variables by size, so that (2) is minimised, but these equations proved troublesome to solve because of dependencies among the components. Since then there have been numerous attempts to obtain efficient approximations to this optimum solution. The first such approximation, suggested by Dalenius and Hodges (1957, 1959), constructs the strata by taking equal intervals on the cumulative function of the square root of the frequencies; this method is still often used today. Eckman's rule (1959) of iteratively equalising the product of stratum weights and stratum ranges was found to require arduous calculations, and is less used than the method of Dalenius and Hodges method (Nicolini 2001). Lavallée and Hidiroglou (1988) derived an iterative procedure for stratifying skewed populations into a take-all stratum and a number of take-some strata such that the sample size is minimised for a given level of reliability. Other recent contributions include Hedlin (2000) who revisited Ekman's rule, Dorfman and Valliant (2000) who compared model-based stratified sampling with balanced sampling, and Rivest (2002) who constructed a generalisation of the Lavallée and Hidiroglou algorithm by providing models accounting for the discrepancy between the stratification variable and the survey variable.

In the present paper we propose an algorithm which is much simpler to implement than any of those currently available. It is based on an observation by Cochran (1961), that with near optimum boundaries the coefficients of variation are often found to be approximately the same in all strata. He concluded however that computing and setting equal the standard deviations of the strata would be too complicated to be feasible in practice. In what follows we show that, for skewed distributions, the coefficients of variation can be approximately equalised between strata

¹ Patricia Gunning, School of Computing, Dublin City University, Dublin 9, Ireland; Jane M. Horgan, School of Computing, Dublin City University, Dublin 9, Ireland.

using the geometric progression. This new algorithm is derived in section 2. Section 3 compares the efficiency of the new approximation with the cumulative root frequency and the Lavallée and Hidirolglou approximations. We summarise our findings in section 4.

2. AN ALTERNATIVE METHOD OF STRATUM CONSTRUCTION

To stratify a population by size is to subdivide it into intervals, with endpoints $k_0 < k_1, < \dots, < k_L$. Ideally, the division should be based on the survey variable Y . Such a construction is of course not possible since Y is unknown; if it were known we would not need to estimate it. In practice therefore we use a known auxiliary variable X , which is correlated with the survey variable.

In order to make the breaks (k_0, k_1, \dots, k_L) for any given k_0 and k_L , we seek to make the $CV_h = S_{xh} / \bar{X}_h$ the same for $h = 1, 2, \dots, L$:

$$\frac{S_{x1}}{\bar{X}_1} = \frac{S_{x2}}{\bar{X}_2} = \dots = \frac{S_{xL}}{\bar{X}_L}. \quad (3)$$

Now S_{xh} is the standard deviation and \bar{X}_h the mean of X in stratum h . If we make the assumption that the distribution within each stratum is approximately uniformly distributed we may write

$$\bar{X}_h \approx \frac{k_h + k_{h-1}}{2}, \quad (4)$$

$$S_{xh} \approx \frac{1}{\sqrt{12}} (k_h - k_{h-1}). \quad (5)$$

As an approximation to the coefficients of variation, this gives

$$CV_h \approx \frac{(k_h - k_{h-1}) / \sqrt{12}}{(k_h + k_{h-1}) / 2} \quad (6)$$

with equal CV_h therefore we must have

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}}. \quad (7)$$

This new and exotic recurrence relation reduces however to something familiar:

$$k_h^2 = k_{h+1} k_{h-1}; \quad (8)$$

the stratum boundaries are the terms of a geometric progression.

$$k_h = ar^h \quad (h = 0, 1, \dots, L). \quad (9)$$

Thus $a = k_0$, the minimum value of the variable, and $ar^L = k_L$, the maximum value of the variable. It follows that the constant ratio can be calculated as $r = (k_L / k_0)^{1/L}$. For a numerical example take

$$L = 4; \quad k_0 = 5; \quad k_4 = 50,000; \quad (10)$$

thus $k_h = 5 \cdot 10^h$ ($h = 0, 1, 2, 3, 4$) and the strata form the ranges

$$5 - 50; 50 - 500; 500 - 5,000; 5,000 - 50,000. \quad (11)$$

This is clearly an extremely simple method of obtaining stratum breaks.

The relationship in (8) depends on the assumption that the distributions within strata are uniform. This may be justified by the following heuristic argument. When the parent distribution is positively skewed, then the low values of the variable have a high incidence, which decreases as the variable values increase, which makes it appropriate to take small intervals at the beginning and large intervals at the end. This is what happens with a geometric series of constant ratio greater than one. In the lower range of the variable, the strata are narrow so that an assumption of rectangular distribution in them is not unreasonable. As the value of the variable increases, the stratum width increases geometrically. This coincides with the decreased rate of change of the incidence of the positively skewed variable, so here also the assumption of uniformity is reasonable.

This algorithm will of course not work for normal distributions. Also since the boundaries increase geometrically, it will not work well with variables that have very low starting points: this will lead to too many small strata; the rule breaks down completely when the lower end point is zero. We expect the best results when the distribution is highly positively skewed and the upper part contains a small percentage of the total frequency.

3. THE PERFORMANCE OF THE ALGORITHM

3.1 Some Real Positively Skewed Populations

To test our algorithm, we implement it on four specific populations, which are skewed with positive tail:

Our first population (Population 1) is an accounting population of debtors in an Irish firm, detailed in Horgan (2003). In addition, we use three of the skewed populations that Cochran (1961) invoked to illustrate the efficiency of

the cumulative root frequency method of stratum construction. These are:

- The population in thousands of US cities (Population 2);
- The number of students in four-year US colleges (Population 3);
- The resources in millions of dollars of a large commercial bank in the US (Population 4).

There were five other populations in the Cochran paper, which turned out to be unsuitable for use with our algorithm. In three cases the variable was a proportion:

agricultural loans, real estate loans and independent loans expressed as a percentage of the total amount of bank loans. Another, a population of farms in which the variable ranged from 1 to 18, was essentially discrete. Yet another, a population of income tax returns, was not sufficiently skewed: it owed its skewness to the top 0.05% of the population, and when this was removed, or put in a take-all stratum, the skewness disappeared.

These four populations are illustrated and summarised in Figure 1 and Table 1 in decreasing order of skewness.

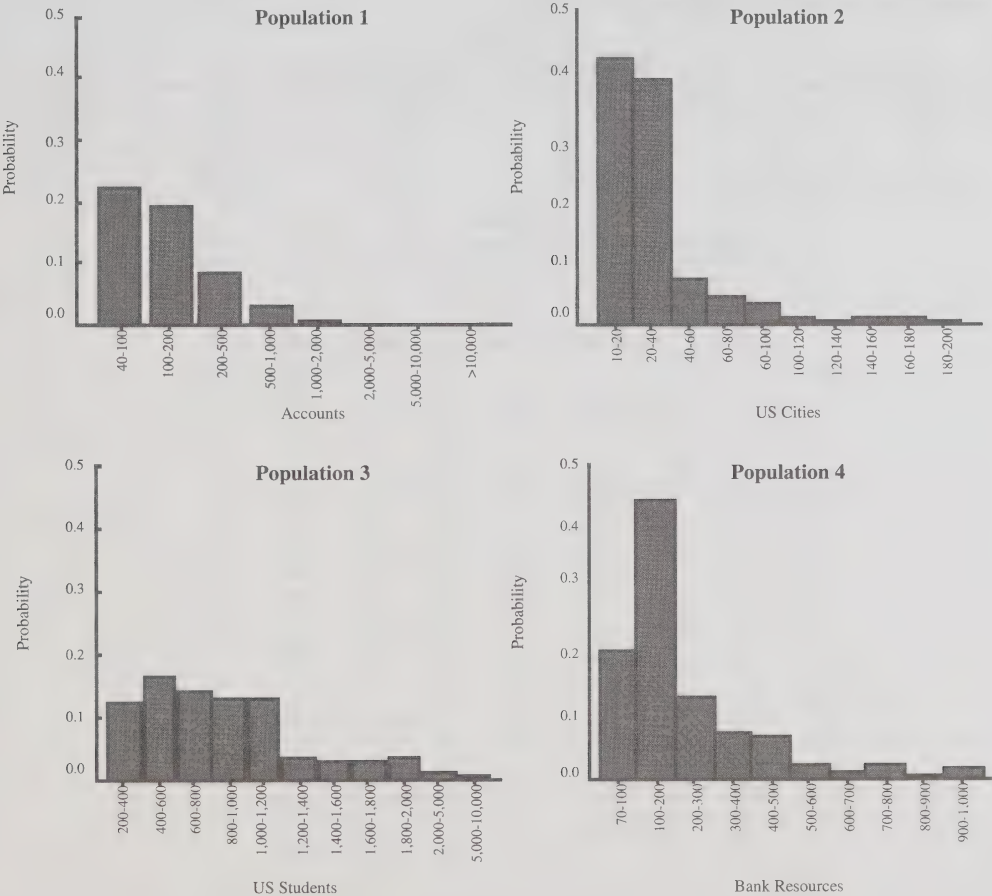


Figure 1. Populations

The new algorithm is implemented on these populations, and compared with the cumulative root frequency (cum \sqrt{f}) and the Lavallée-Hidiroglou methods of stratum construction.

3.2 Comparison with the Cumulative Root Frequency Method

We first compare the performance of the new algorithm with cum \sqrt{f} by dividing the populations summarised in Table 1 into $L = 3, 4$ and 5 strata, using both methods to make the breaks. The results are given in Tables 2, 3 and 4.

A cursory examination of the coefficients of variation in Tables 2, 3 and 4 suggests that, in most cases, the geometric method is more successful than cum \sqrt{f} in obtaining near-equal strata CV_h . For example in Population 1, which has the greatest skewness, the CV_h differ substantially from

each other when cum \sqrt{f} is used to make the breaks, while the geometric method appears to achieve near-equal CV_h in all cases of 3, 4 and 5 strata: the best results are obtained with $L = 5$. In the other three populations, the CV_h are not as diverse with cum \sqrt{f} , but they still appear more variable than those obtained with the geometric method of stratum construction.

The CV_h with the geometric method are more homogeneous when $L = 4$ or 5 than when $L = 3$; this is to be expected since the validity of the assumption of uniformity of the distribution of elements within stratum is strengthened with increased number of strata.

A more detailed analysis of the variability of the CV_h between strata is given in Table 5, where the standard deviation of the CV_h is calculated for each design.

Table 1
Summary Statistics for Real Populations

Population	N	Range	Skewness	Mean	Variance
1	3,369	40 – 28,000	6.44	838.64	3,511,827
2	1,038	10 – 200	2.88	32.57	924
3	677	200 – 10,000	2.46	1,563.00	3,236,602
4	357	70 – 1,000	2.08	225.62	36,274

Table 2
The Geometric vs the Cum \sqrt{f} : Stratum Breaks with $L = 3$ and $n = 100$

Population	Stratification Method	CV	Stratum		
			1	2	3
1	Geometric	0.0600	k_h 354	3,152	
			N_h 2,334	1,288	189
			n_h 9	46	45
			CV_h 0.71	0.68	0.64
			k_h 558	2,236	
			N_h 2,339	735	295
	Cum \sqrt{f}	0.0600	n_h 19	17	64
			CV_h 0.70	0.42	0.76
	Geometric	0.0270	k_h 26	72	
			N_h 701	243	94
			n_h 36	29	35
			CV_h 0.28	0.23	0.33
2	Cum \sqrt{f}	0.0269	k_h 28	66	
			N_h 729	208	101
			n_h 40	22	38
			CV_h 0.29	0.25	0.34
	Geometric	0.0317	k_h 726	2,645	
			N_h 253	321	103
3	Cum \sqrt{f}	0.0282	n_h 9	38	53
			CV_h 0.32	0.37	0.39
			k_h 1,179	3,629	
			N_h 456	152	69
			n_h 37	35	28
			CV_h 0.41	0.31	0.27
4	Geometric	0.0184	k_h 168	405	
			N_h 211	93	53
			n_h 27	27	46
			CV_h 0.23	0.24	0.30
	Cum \sqrt{f}	0.0198	k_h 162	441	
			N_h 207	107	43
			n_h 25	39	36
			CV_h 0.23	0.30	0.27

Table 3
The Geometric vs the Cum \sqrt{f} : Stratum Breaks with $L = 4$ and $n = 100$

Population	Stratification	CV	Stratum					
	Method		1	2	3	4		
1	Geometric	0.0430	k_h	205	1,057	5,443		
			N_h	1,416	1,382	483	88	
			n_h	6	22	40	32	
	Cum \sqrt{f}	0.0480	CV_h	0.45	0.44	0.48	0.50	
			k_h	558	1,117	2,795		
			N_h	2,339	483	325	222	
			n_h	23	5	10	62	
			CV_h	0.70	0.19	0.27	0.69	
			k_h	20	43	93	200	
	2	Geometric	0.0194	N_h	459	398	130	51
n_h				22	31	25	22	
CV_h				0.22	0.20	0.22	0.22	
Cum \sqrt{f}		0.0213	k_h	19	38	85		
			N_h	393	428	155	62	
			n_h	15	26	30	29	
			CV_h	0.20	0.17	0.25	0.26	
			k_h	526	1,386	3,653		
			N_h	138	343	127	69	
3		Geometric	0.0214	n_h	5	27	26	42
	CV_h			0.27	0.26	0.26	0.27	
	k_h			690	2,160	5,100		
	Cum \sqrt{f}	0.0230	N_h	235	319	75	48	
			n_h	13	43	21	23	
			CV_h	0.31	0.33	0.29	0.19	
	4	Geometric	0.0142	k_h	134	261	504	
				N_h	156	109	63	29
				n_h	20	23	29	28
		Cum \sqrt{f}	0.0143	CV_h	0.18	0.19	0.19	0.20
k_h				162	255	488		
N_h				207	58	57	35	
			n_h	33	9	23	35	
			CV_h	0.23	0.11	0.18	0.24	
			k_h					

Table 4
The Geometric vs the Cum \sqrt{f} : Stratum Breaks with $L = 5$ and $n = 100$

Population	Stratification Method	CV	Stratum					
			1	2	3	4	5	
1	Geometric	0.0360	k_h	147	549	2,037	7,552	
			N_h	1,054	1,267	732	265	51
			n_h	2	14	27	33	24
			CV_h	0.37	0.38	0.40	0.37	0.41
			k_h	279	838	1,677	4,193	
	Cum \sqrt{f}	0.0349	N_h	1,644	1,010	332	249	134
			n_h	9	14	7	15	55
			CV_h	0.52	0.30	0.20	0.25	0.57
			k_h	17	32	59	108	
			N_h	364	418	130	87	39
2	Geometric	0.0144	n_h	18	28	17	20	17
			CV_h	0.18	0.14	0.15	0.16	0.15
			k_h	28	38	57	104	
			N_h	729	92	89	88	40
			n_h	58	4	7	16	15
	Cum \sqrt{f}	0.0186	CV_h	0.28	0.08	0.11	0.16	0.16
			k_h	433	941	2,043	4,434	
			N_h	100	255	1,989	74	56
			n_h	2	16	27	20	35
			CV_h	0.22	0.21	0.24	0.21	0.21
3	Geometric	0.0184	k_h	1,179	1,669	3,139	6,079	
			N_h	50	3	17	15	15
			n_h	0.40	0.09	0.20	0.19	0.13
			CV_h	118	200	339	576	
			N_h	114	116	64	39	24
	Cum \sqrt{f}	0.0212	n_h	12	20	24	18	24
			CV_h	0.14	0.14	0.17	0.12	0.16
			k_h	162	255	395	627	
			N_h	207	58	37	36	19
			n_h	44	11	10	19	16
4	Geometric	0.0110	CV_h	0.23	0.11	0.10	0.13	0.11
			k_h					
			N_h					
			n_h					
			CV_h					
	Cum \sqrt{f}	0.0119	k_h					
			N_h					
			n_h					
			CV_h					
			k_h					

Table 5
The Variability of the CV_h for the Geometric and the Cum \sqrt{f} Methods

Strata		Population			
		1	2	3	4
3	Geometric	0.035	0.050	0.036	0.038
	Cum \sqrt{f}	0.181	0.045	0.072	0.035
4	Geometric	0.027	0.010	0.006	0.008
	Cum \sqrt{f}	0.276	0.042	0.062	0.059
5	Geometric	0.018	0.015	0.013	0.020
	Cum \sqrt{f}	0.166	0.076	0.119	0.054

We see from Table 5 that, with just two exceptions, the standard deviations of the CV_h are substantially lower with the geometric method of stratum construction than with cum \sqrt{f} . In the two cases where the cumulative root has a lower standard deviation than the geometric, the differences between them is not great, and occur with the smallest number of strata, $L=3$, in Populations 2 and 4. We may conclude therefore that the new algorithm is successful in breaking the strata in such a way that the CV_h are near equal.

What remains is to investigate whether the geometric breaks lead to more efficient estimation than cum \sqrt{f} . To do this, the two methods are compared in terms of the relative efficiency or variance ratio obtained with $n = 100$ allocated optimally among the strata using *Neyman allocation* (Neyman 1934):

$$n_h = \left(\frac{N_h S_{xh}}{\sum_{i=1}^L N_i S_{xi}} \right) n. \quad (12)$$

The relative efficiency is defined as

$$eff_{\text{cum, geom}} = \frac{V_{\text{cum}}(\bar{x}_{st})}{V_{\text{geom}}(\bar{x}_{st})}, \quad (13)$$

where $V_{\text{cum}}(\bar{x}_{st})$ and $V_{\text{geom}}(\bar{x}_{st})$ are the variances of the mean respectively with the cumulative root frequency and the geometric methods, with $n = 100$ and n_h allocated as in (12) for each of the stratification methods. In sample size planning the relative efficiencies may be interpreted as the proportionate increase or decrease in the sample size with cum \sqrt{f} to obtain the same precision as that of the geometric method with $n = 100$.

The variance calculations are based on the auxiliary variable X , and since this is assumed to be highly correlated with the unknown survey variable Y , we can assume the relative efficiency eff , given in (13), will be a reasonable approximation of the relative efficiency of Y .

Table 6 gives the variance ratio when the number of strata $L = 3, 4$ and 5.

From Table 6 we see that, while this new method is not always more efficient than the cumulative root frequency method of stratum construction, when it is, it is substantially

so, and when it is not it is only marginally worse. For example, large gains in efficiency are observed when $L = 5$ in Populations 2, 3 and 4; here the relative efficiencies are 1.69, 1.33 and 1.17 respectively indicating that samples of sizes $n = 169, 133$ and 117 are required with cum \sqrt{f} to obtain the sample precision as that of the geometric method with $n = 100$.

Table 6
Efficiencies of the Cum \sqrt{f} Relative to the Geometric Method

Strata	Population			
	1	2	3	4
3	0.97	0.99	0.79	1.16
4	1.23	1.19	1.16	1.04
5	0.94	1.69	1.33	1.17

We also see from Table 6 that while there are four cases where the relative efficiency is less than 1, with one exception, all are greater than 0.9. The exception is Population 3 with $L = 3$, the smallest number of strata; the relative efficiency in this case is 0.79.

3.3 Comparison with the Lavallée and Hidirolglou Algorithm

With the Lavallée-Hidirolglou algorithm, the optimum boundaries $k_1, k_2 \dots k_{L-1}$ are chosen to minimise the sample size n for a given level of precision. The requirement on precision is usually stated by requiring the coefficient of variation to be equal to some specified level between 1% – 10%. Obtaining the minimum n is an iterative process, and the SAS code used for implementing it was obtained from the web at <http://www.ulval.ca/pages/lpr/>.

To compare the performance of the new method with Lavallée-Hidirolglou, the CVs from the geometric algorithm given in Tables 2, 3 and 4 are used as input for the Lavallée-Hidirolglou algorithm, and the sample sizes required to obtain the same precision as that of the geometric method with $n = 100$ are computed. The results are given in Table 7.

The first thing to notice from Table 7 is that the sample size required with the Lavallée-Hidirolglou algorithm to obtain the same precision as the geometric method is greater than 100 in all but four cases. In Population 2 with 5 strata, it is necessary to increase the sample size by 36% to

$n = 136$, to obtain the same precision as the geometric method with $n = 100$. With three and four strata, sample sizes of $n = 121$ and 113 are required in Population 1, and samples sizes of $n = 123$ and $n = 117$ are required in Population 2, to obtain the same precision as the geometric method. When the sample size falls below $n = 100$, the drop is not as large. In Population 4, with four and five strata, $n = 93$ and $n = 99$ respectively, and in Population 1 with 5 strata a sample size of $n = 90$ will suffice with the Lavallée-Hidiroglou algorithm to obtain the same precision as the geometric method.

The results in Table 7 might appear to indicate that the geometric method outperforms the Lavallée-Hidiroglou

method in terms of the minimum sample size required for a specified precision. We observe however that the geometric method does not give a take-all stratum. If this is required it is more appropriate to use the Lavallée-Hidiroglou to obtain the strata. Often, in financial applications the top stratum is decided judgementally; for example US state taxing authorities typically decide their take-all stratum based on a total percentage of purchase amounts (Falk, Rotz and Young 2003). If after such a take-all stratum has been removed the skewness remains, the geometric method is probably the easier and more efficient way of obtaining the remaining strata.

Table 7
Boundaries and Sample Size Required with the Lavallée-Hidiroglou Method to Obtain the Same CV as the Geometric Method when $n = 100$

Population	n	CV	3 Strata					
			1	2	3			
1	121	0.0600	k_h	1,248	8,676			
			N_h	2,867	464	38		
			n_h	42	41	38		
			CV_h	0.87	0.57	0.37		
2	123	0.0270	k_h	35	102			
			N_h	795	202	41		
			n_h	47	35	41		
			CV_h	0.31	0.31	0.17		
3	107	0.0317	k_h	1,398	4,197			
			N_h	481	135	61		
			n_h	28	18	61		
			CV_h	0.41	0.30	0.24		
4	100	0.0184	k_h	172	361			
			N_h	212	85	60		
			n_h	22	18	60		
			CV_h	0.23	0.21	0.32		
			4 Strata					
			1	2	3	4		
1	113	0.0430	k_h	442	1,828	8,411		
			N_h	2,086	915	327	41	
			n_h	16	21	35	41	
			CV_h	0.64	0.41	0.45	38	
2	117	0.0194	k_h	19	37	95		
			N_h	393	420	176	49	
			n_h	13	21	34	49	
			CV_h	0.19	0.16	0.28	0.21	
3	103	0.0214	k_h	740	1,505	3,819		
			N_h	256	234	118	69	
			n_h	9	10	15	69	
			CV_h	0.32	0.18	0.25	0.27	
4	93	0.0142	k_h	117	188	359		
			N_h	111	112	74	60	
			n_h	7	9	17	60	
			CV_h	0.14	0.12	0.19	0.32	
			5 Strata					
			1	2	3	4	5	
1	90	0.0360	k_h	342	1,153	3,431	10,301	
			N_h	1,846	993	357	147	26
			n_h	12	14	17	21	26
			CV_h	0.58	0.34	0.31	0.31	0.32
2	136	0.0144	k_h	14	21	35	80	
			N_h	189	270	336	164	79
			n_h	4	7	16	30	79
			CV_h	0.12	0.10	0.12	0.24	0.30
3	105	0.0184	k_h	512	869	1,577	3,675	
			N_h	133	180	185	110	69
			n_h	4	5	10	17	69
			CV_h	0.27	0.15	0.16	0.23	0.27
4	99	0.0119	k_h	99	130	189	339	
			N_h	70	68	85	71	63
			n_h	4	4	8	20	63
			CV_h	0.10	0.08	0.10	0.18	0.33

4. SUMMARY

This paper derives a simple algorithm for the construction of stratum boundaries in positively skewed populations, for which it is shown that the stratum breaks may be obtained using the geometric distribution. The proposed method is easier to implement than approximations previously proposed. Comparisons with the commonly used cumulative root frequency method using four positively skewed real populations divided into three, four and five strata, showed substantial gains in the precision of the estimator of the mean; the greatest gains occurring when the number of strata was five. Comparisons with the Lavallée-Hidiroglou method indicated that a greater sample size was required to obtain the same precision as the geometric method in most cases; the greatest increase in the required sample size occurred with the largest number of strata. One limitation of the new algorithm compared to the Lavallée-Hidiroglou method of stratum construction is that it does not determine a take-all top stratum.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Irish Research Council for Science, Engineering and Technology.

We are indebted to the referees for their helpful suggestions which have greatly improved the original paper.

REFERENCES

COCHRAN, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 2, 345-358.

- DALENIUS, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 203-213.
- DALENIUS, T., and HODGES, J.L. (1957). The choice of stratification points. *Skandinavisk Aktuarietidskrift*, 198-203.
- DALENIUS, T., and HODGES, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 88-101.
- DORFMAN, A.H., and VALLIANT, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.
- ECKMAN, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 219-229.
- FALK, E., ROTZ, W. and YOUNG, L.L.P. (2003). Stratified sampling for sales and use tax highly skewed data-determination of the certainty stratum cut-off amount. *Proceedings of the Section on Statistical Computing*, American Statistical Association, 66-72.
- HEDLIN, D. (2000). A procedure for stratification by an extended ekman rule. *Journal of Official Statistics*, 16, 15-29.
- HORGAN, J.M. (2003). A list sequential sampling scheme with applications in financial auditing. *IMA Journal of Management Mathematics*, 14, 1-18.
- LAVALLÉE, P., and HIDIROGLOU, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.
- NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistics Society*, 97, 558-606.
- NICOLINI, G. (2001). A method to define strata boundaries. Working Paper 01-2001-marzo, Dipartimento di Economia Politica e Aziendale, Università degli Studi di Milano.
- RIVEST, L.-P. (2002). A generalization of the Lavallée-Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198.

Feeding Back Information on Ineligibility from Sample Surveys to the Frame

DAN HEDLIN and SUOJIN WANG¹

ABSTRACT

It is usually discovered in the data collection phase of a survey that some units in the sample are ineligible even if the frame information has indicated otherwise. For example, in many business surveys a nonnegligible proportion of the sampled units will have ceased trading since the latest update of the frame. This information may be fed back to the frame and used in subsequent surveys, thereby making forthcoming samples more efficient by avoiding sampling ineligible units. On the first of two survey occasions, we assume that all ineligible units in the sample (or set of samples) are detected and excluded from the frame. On the second occasion, a subsample of the eligible part is observed again. The subsample may be augmented with a fresh sample that will contain both eligible and ineligible units. We investigate what effect on survey estimation the process of feeding back information on ineligibility may have, and derive an expression for the bias that can occur as a result of feeding back. The focus is on estimation of the total using the common expansion estimator. An estimator that is nearly unbiased in the presence of feed back is obtained. This estimator relies on consistent estimates of the number of eligible and ineligible units in the population being available.

KEY WORDS: Dead unit; Feed back bias; Overcoverage; Permanent random number sampling; Panel survey; Coordinated samples.

1. INTRODUCTION

To facilitate estimation of change, consecutive samples in a repeated survey are usually overlapping. If several surveys draw samples from the same frame, it is often desirable to spread the response burden out by making sure that samples for different surveys are not overlapping to a greater extent than necessary. This is particularly desirable if the frame is moderately large and used for many continuing surveys, which is a situation that many national statistical institutes face when conducting business surveys. Stratified simple random sampling is a very common design for business surveys. The skewed distribution of businesses calls for large sampling fractions in many strata, which aggravates the response burden for medium size and large businesses. Both estimation of change and response burden issues are of paramount importance in official business statistics. Therefore, sampling systems have been constructed that allow the organisation to co-ordinate samples, either positively or negatively (*i.e.* to create overlap or to make sure that there is little overlap).

For example, the Office for National Statistics (ONS) in the United Kingdom uses the Permanent Random Number (PRN) technique, which is a widely used method for drawing samples from lists. A PRN from the uniform distribution on $[0,1]$ is attached to each frame unit independently of each other and independently of the unit labels and any variables associated with the units. Each unit will retain the

PRN throughout its existence. The units can be ordered along a line starting at 0 and ending at 1 and we refer to this line as the *PRN line*. To draw a simple random sample without replacement, an S_i , with a predetermined sample size n , a point is selected (randomly or purposively) on the PRN line and the n units to the right (say) are included in the sample. Two SIs are fully co-ordinated if they are drawn from the same interval. For overviews and further details see Ohlsson (1995) and Ernst, Valliant and Casady (2000).

Samples for repeated surveys can also be selected with a panel technique where a set of rotation groups are selected at the first wave and one, say, of the groups is replaced with a fresh rotation group at the second wave and the other groups are retained in the sample. The difference between PRN sampling and panel sampling is more about the way to control overlaps than having different sampling designs.

There are in principle two main sources of data that are used to maintain a frame: administrative ones and surveys. Various administrative bodies send tapes to the ONS on a regular basis with information on, *e.g.*, births and deaths of businesses. While these tapes are sent to the ONS very frequently, the distribution of the time it takes for a new unit or an alteration of an old unit to be registered on the frame is highly skewed. This is partly due to frame maintenance procedures, *e.g.* to avoid duplicates. There is also very often a considerable difference in time between the actual and formal termination of a business. Therefore, most of the ONS's business surveys share the information on deaths

¹ Dan Hedlin, Statistics Sweden, Box 24 300, SE-104 51 Stockholm, Sweden. E-mail: dan.hedlin@scb.se; Suojin Wang, Texas A&M University, Department of Statistics, College Station, Texas 77843-3143, U.S.A. E-mail: sjwang@stat.tamu.edu.

they obtain through their samples with other business surveys to speed up the information process. We examine the effects of using sample surveys to update a frame that is used for repeated surveys. This is in principle how information on dead units is treated in business surveys at the ONS, Statistics Sweden, and some other national statistical institutes.

It would seem natural that this new information should be made available to other sample surveys, which otherwise may include the dead units in their samples and therefore lose precision. However, as pointed out by Srinath (1987) among others, such a procedure may cause bias. We refer to this as *feed back bias*, which results whenever the sampling mechanism is not independent of the feed back procedure. For example, consider a situation where all dead units are found and deleted at the first wave of a panel survey. If no further deaths have occurred up to the second-wave observation of the panel units, the second-wave sample contains only live units. Without knowledge of the total number of live units in the population at the time of the second wave, an unbiased estimator of the total cannot be constructed. While more information about the population has been gathered when the deaths were recorded at the first wave, there is actually less information in the second wave-sample on the proportion of live units in the population. We show how an estimate of the number of live units in the population can be used to construct an approximately unbiased estimate of the population total.

A safe recommendation would be that no information on deaths from sample surveys, other than from completely enumerated strata, may be used to update the frame when samples are co-ordinated over time (cf. Ohlsson 1995, page 168, and Colledge 1989, page 103). However, to prohibit feeding back seems to deny oneself the use of all available information. We obtain an expression for the feed back bias and show that the feed back bias can be estimated and used to adjust conventional estimators. Schioppa-Kratina and Srinath (1991) adjust the sampling weights to counter an expected too low proportion of dead units in the rotating sample of the Survey of Employment, Payroll and Hours conducted by Statistics Canada. Hidioglou and Laniel (2001) discuss the feed back issue briefly. A general discussion of frame issues is given by Colledge (1995) and overviews of issues associated with continuing business surveys include College (1989), Hidioglou and Srinath (1993), Srinath and Carpenter (1995), and Hidioglou and Laniel (2001).

Instead of the terms eligible and ineligible we use the more emotive words dead and live, although our reasoning does cover all kinds of ineligibility. The discussion is confined to the estimation of the total

$$t_y = \sum_U y_k \quad (1)$$

of some study variable $\mathbf{y}' = (y_1, y_2, \dots, y_N)$ on a population U with unit labels $\{1, 2, \dots, N\}$.

When the sampled units are observed, we assume that all dead units in the sample are classified as dead and the frame is updated with this information. This may be difficult in practice. In some surveys, however, the eligibility of all nonresponding units can be correctly identified.

Section 2 introduces the necessary notation and concepts and gives expressions for the feed back bias when estimating a total. Section 3 discusses three strategies that may be used in the presence of feed back and compares these in a simulation study. The paper concludes with a discussion in section 4.

2. EXPRESSIONS FOR FEED BACK BIAS

2.1 Introduction and Notation

We assume throughout that a dead unit is always out of scope and that the value of the study variable of a dead unit is always zero. (It is conceivable that dead units are eligible in some surveys; for example, a business survey collecting data on production may have defined businesses that were alive at least part of the reference period as eligible.) We adopt the design-based view that the survey population and the study variable are fixed and non-stochastic at any given point in time. The situation we address is as follows. One or more samples are drawn from the frame which comprises the original survey population, U_1 . Let the set of samples drawn from U_1 be denoted by s_1 . For convenience we assume that the frame units and population units are of the same type. We refer to the updated frame, where all dead units that have been included in samples from U_1 have been excluded, as the current survey population, U_2 . For example, two surveys may simultaneously work with a sample each, and after they have fed back, U_1 has shrunk to U_2 . We disregard births of new units and other deaths than those deleted through samples from U_1 . We will also disregard undercoverage, nonresponse and measurement errors. In practice, administrative sources will provide information on deaths. They work independently from the sampling procedures employed by the statistical agency and will therefore not contribute to feed back bias. These units are dead by administrative sources. We can think of these dead units as being excluded from the population. See Hidioglou and Laniel (2001) for a discussion of estimation in the presence of units deceased by administrative sources. While the sampling design here is assumed to be SI, it can readily be extended to stratified simple random sampling.

Let $U_{2,d}$ and $U_{2,l}$ be the two subsets of the current survey population, $U_2 = U_{2,d} \cup U_{2,l}$, that consist of dead and live units, respectively. All units in $U_{2,d}$ and $U_{2,l}$ are assumed to be flagged as live. Units that are flagged as dead but for which the independence of detection and the sampling mechanism cannot be assured are called *dead by sample survey sources*. In our set-up, these are the dead units detected in samples taken from U_1 . Let the set of these units be denoted by $s_{1,d}$, and we have the relationship $U_1 = U_2 \cup s_{1,d}$. Figure 1 displays the sets and their relationships. Let N and n with a proper subscript be the size of the corresponding population and sample(s), respectively. Then $N_1 = N_2 + n_{1,d}$ and $N_2 = N_{2,l} + N_{2,d}$. At the time when samples are drawn from U_2 , N_2 and $n_{1,d}$ are known numbers, whereas $N_{2,l}$ and $N_{2,d}$ are unknown. Moreover, $n_{1,d}$, $N_{2,d}$ and N_2 could be viewed as random depending on feed back results, while $N_{2,l}$ is fixed. Following principles of Durbin (1969) and more recently in Thompson (1997), we would in many situations prefer to condition on $n_{1,d}$. For example, if it is seen that $n_{1,d} = 0$, then it does not seem appropriate to include in the inference the possibility that $n_{1,d}$ could have been large. However, to analyse the development of the feed back bias over a series of waves in a panel survey when planning the survey, unconditional analysis would be preferable. We also provide an expression for the unconditional feed back bias.

Denote by $s_{1,l}$ the live part of s_1 , i.e., the part of U_2 that was covered by the previous sample(s) drawn from U_1 ; see Figure 1. Clearly, $s_{1,l}$ is a random set and we have $s_{1,l} \subset U_{2,l}$. Let the nonsampled part of U_2 be denoted by $U_{2,wd}$ ('wd' for 'with dead units'). It is also a random set and

encompasses all of $U_{2,d}$ and part of $U_{2,l}$. We have $U_2 = U_{2,wd} \cup s_{1,l}$.

Let s_2 be an SI taken from U_2 . Estimators based on s_2 will suffer from feed back bias unless special information is at hand, such as knowledge about $N_{2,l}$, which is not usually the case. To derive an expression for the feed back bias we shall first obtain the inclusion probabilities. To do this, it is useful to consider the two sample parts of s_2 separately: the sample part $s_{2,a}$ of size $n_{2,a}$ taken from $s_{1,l}$ through PRN sampling or a panel sampling technique, and the remaining part $s_{2,b}$ taken from $U_{2,wd}$. If the sampling is done with a panel technique, the sample parts $s_{2,a}$ and $s_{2,b}$ are the old and new rotation groups, respectively. If the sample is drawn with PRN sampling, $s_{2,a}$ and $s_{2,b}$ consist of units with PRN's that fell in s_1 or did not fall in s_1 , respectively. Whether the sample was drawn through PRN sampling or a panel sampling technique, the sample parts can be viewed as two fixed size samples, each drawn with the SI design from their respective subpopulation. We condition on $n_{2,a}$ and $n_{2,b}$ throughout without making it explicit in formulae. With the notation $(k \in s_{2,a})$ we refer to the event that a unit is first included in the first-wave sample(s) from U_1 and then in the second-wave sample taken from what remains of the first-wave sample(s) after dead units have been taken out. The notation $(k \in s_{2,b})$ is analogous. Let $I(k \in s_{2,a}) = 1$ when unit k is included in $s_{2,a}$, otherwise $I(k \in s_{2,a}) = 0$. To derive the overall bias it is convenient to analyse the biases from the sample parts $s_{2,a}$ and $s_{2,b}$. We derive an expression for each of these in section 2.2 and section 2.3, respectively, and in section 2.4 the bias expressions will be amalgamated.

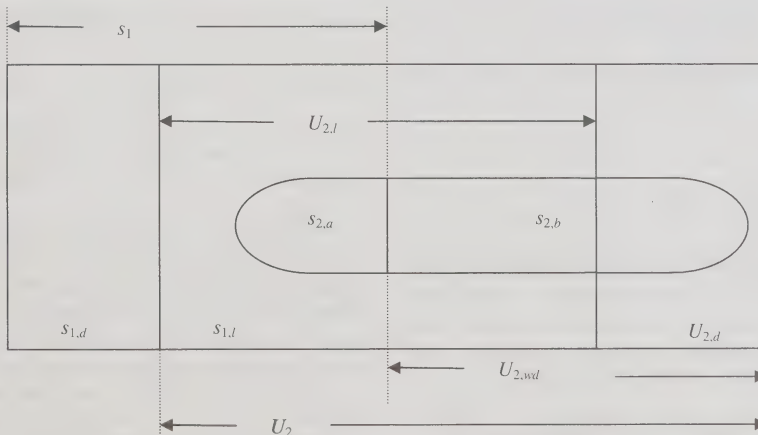


Figure 1. The original survey population, U_1 , and its subsets. The grey area represents s_2 , the sample from U_2 .

2.2 Feed Back Bias from a Sub-sample from the Original Sample

Suppose a sub-sample $s_{2,a}$ is taken from $s_{1,d}$, the live part of the first-wave sample(s). Recall that $y_k = 0$ if k is a dead unit and that $U_2 = U_{2,d} \cup U_{2,l}$. Thus we have $\sum_{s_{2,a}} y_k = \sum_{U_{2,l}} y_k I(k \in s_{2,a}) = \sum_{U_2} y_k I(k \in s_{2,a})$. Assume that $N_{2,l} > 0$. Then we obtain that $\Pr[k \in s_{2,a} | n_{1,d}] = n_{2,a} / N_{2,l}$, since a sample of size $n_{2,a}$ is effectively selected from a population of size $N_{2,l}$ with the SI design (through an SI sample from U_1 followed by an SI sample from $U_{2,l}$). Note that a unit k in $s_{2,a}$ must be alive since $U_{2,l}$ consists solely of live units.

Denote the bias of an estimator $\hat{\theta}$ for the parameter θ by $B(\hat{\theta}, \theta)$. Then with respect to the population total $t_y = \sum_{U_2} y_k$, the conditional bias of a general linear estimator $\hat{t}_y^{(s_{2,a})} = \sum_{s_{2,a}} w_k y_k$ based on $s_{2,a}$, with any given w_k 's, is

$$\begin{aligned} B(\hat{t}_y^{(s_{2,a})}, t_y | n_{1,d}) &= \sum_{U_{2,l}} \{w_k \Pr[k \in s_{2,a} | n_{1,d}] - 1\} y_k \\ &= \sum_{U_{2,l}} \left(\frac{w_k n_{2,a}}{N_{2,l}} - 1 \right) y_k \\ &= \sum_{U_2} \left(\frac{w_k n_{2,a}}{N_{2,l}} - 1 \right) y_k. \end{aligned} \quad (2)$$

For the sample part $s_{2,a}$, the naive expansion estimator that ignores feed back bias would have weights $w_k = N_2 / n_{2,a}$. From (2) the bias of the estimator $\hat{t}_{y\pi}^{(s_{2,a})} = N_2 / n_{2,a} \sum_{s_{2,a}} y_k$ is

$$B(\hat{t}_{y\pi}^{(s_{2,a})}, t_y | n_{1,d}) = \frac{N_{2,d}}{N_{2,l}} t_y. \quad (3)$$

2.3 Feed Back Bias from a Sample Taken Afresh from the Current Survey Population

Next, we derive the bias arising from the sample part $s_{2,b}$ of size $n_{2,b}$ taken from U_2 through $U_{2,wd}$, see Figure 1. First note that

$$\Pr[k \in s_{2,b} | k \in U_{2,wd}, n_{1,d}] = \frac{n_{2,b}}{N_{2,wd}}. \quad (4)$$

From (4) we obtain that the conditional expected value of $\hat{t}_y^{(s_{2,b})} = \sum_{s_{2,b}} w_k y_k$ is

$$\begin{aligned} E(\hat{t}_y^{(s_{2,b})} | n_{1,d}) &= E \left[\frac{n_{2,b}}{N_{2,wd}} \sum_{U_{2,wd}} w_k y_k | n_{1,d} \right] \\ &= \frac{n_{2,b}}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} \sum_{U_2} w_k y_k. \end{aligned}$$

The second equation above is due to the fact that given $n_{1,d}$, all $N_{2,l}$ live units in U_2 are equally likely to be in

$U_{2,wd}$, which has $N_{2,l} - n_{1,l}$ live units. Therefore, the conditional bias of $\hat{t}_y^{(s_{2,b})}$ is

$$B(\hat{t}_y^{(s_{2,b})}, t_y | n_{1,d}) = \sum_{U_2} \left(\frac{w_k n_{2,b}}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} - 1 \right) y_k. \quad (5)$$

For the expansion estimator $\hat{t}_{y\pi}^{(s_{2,b})}$ with weights $w_k = N_2 / n_{2,b}$ the bias is

$$B(\hat{t}_{y\pi}^{(s_{2,b})}, t_y | n_{1,d}) = B t_y, \quad (6)$$

where

$$\begin{aligned} B &= \frac{N_2}{N_{2,wd}} \frac{N_{2,l} - n_{1,l}}{N_{2,l}} - 1 \\ &= \frac{N_2(N_{2,l} - n_{1,l}) - N_{2,l}(N_2 - n_{1,l})}{N_{2,wd} N_{2,l}} \\ &= -\frac{N_{2,d} n_{1,l}}{N_{2,l} N_{2,wd}} \\ &= -\frac{N_{2,d}(n_1 - n_{1,d})}{N_{2,l}(N_1 - n_1)}. \end{aligned}$$

The bias is always non-positive since $B \leq 0$. It is easy to see that B is an increasing function of $n_{1,d}$ since $N_{2,d} = N_{1,d} - n_{1,d}$, where $N_{1,d}$ is the fixed number of all dead units in U_1 . It is also readily seen that the maximum of B is attained when $s_{1,d}$ encompasses all dead units in U_1 , that is, when $n_{1,d} = N_{1,d}$ and consequently $N_{2,d} = 0$.

2.4 Feed Back Bias from Sample Parts Combined

Combining (6) with (3) we obtain the overall bias of $\hat{t}_{y\pi} = N_2 / n_2 \sum_{s_2} y_k$ to be

$$\begin{aligned} B(\hat{t}_{y\pi}, t_y | n_{1,d}) &= E(\hat{t}_{y\pi} | n_{1,d}) - t_y \\ &= \frac{N_{2,d}}{N_{2,l}} \left(\frac{n_{2,a}}{n_2} - \frac{n_{2,b}}{n_2} \frac{n_{1,l}}{N_{2,wd}} \right) t_y = \tilde{c} t_y. \end{aligned} \quad (7)$$

The bias in the expansion estimator is really down to not knowing the correct population size. In (3) the bias stems from multiplying the sample average over live units with N_2 rather than the unknown $N_{2,l}$. The bias from the sample parts $s_{2,a}$ and $s_{2,b}$ will in absolute terms be less than (3) and (6), respectively, if some of the dead units in the samples from U_1 have not been identified as dead and therefore have not been weeded out. This would happen, for example, if the status of nonresponding units is difficult to determine.

An unconditional analysis in the presence of feed back can be obtained directly by taking expectation of (7) with respect to $n_{1,d}$. Thus, unconditionally, we have

$$\begin{aligned}
& E\left(\frac{N_2}{n_2} \sum_{s_2} y_k\right) - t_y \\
&= \left[\frac{N_{1,d} - E(n_{1,d})}{N_{2,l}} \left(\frac{n_{2,a}}{n_2} - \frac{n_{2,b}}{n_2} \frac{n_1 - E(n_{1,d})}{N_{2,wd}} \right) \right. \\
&\quad \left. - \frac{n_{2,b}}{n_2 N_{2,l} N_{2,wd}} V(n_{1,d}) \right] t_y \\
&= ct_y,
\end{aligned} \tag{8}$$

where $E(n_{1,d}) = n_1 N_{1,d} / N_1$ and $V(n_{1,d}) = n_1 N_{1,d} N_{2,l} / N_1^2$.

Lavallée (1996) took an interesting approach to a similar problem with panel survey data. In that paper, the problem of frame update using panels with rotation is addressed among other issues. Our approach is different from the approach of that paper in that we consider the two conditional probabilities $\Pr[k \in s_{2,a} | n_{1,d}]$ and $\Pr[k \in s_{2,b} | n_{1,d}]$ separately.

3. THREE SIMPLE STRATEGIES AND A SIMULATION STUDY

3.1 Strategies in the Presence of Feed Back

A strategy, which is referred to as Strategy 1 here, is to feed back, delete the set $s_{1,d}$ from the frame and accept the feed back bias. However, the size of the bias is seldom known. The estimator for Strategy 1 under SI is $\hat{t}_{y\pi} = N_2 / n_2 \sum_{s_2} y_k$ where s_2 is a sample taken from U_2 . To obtain Strategy 2, note that if consistent estimates of $N_{2,d}$ and $N_{2,l}$ are available these may be plugged into (7) or (8) and an estimator with favourable properties is obtained:

$$\hat{t}'_{y\pi} = \hat{t}_{y\pi} (1 + \hat{c})^{-1}, \tag{9}$$

where

$\hat{c} = (\hat{N}_{2,d} / \hat{N}_{2,l}) [n_{2,a} / n_2 - \{n_{2,b} (n_1 - n_{1,d})\} / \{n_2 (N_1 - n_1)\}]$ for both the conditional and unconditional cases since the term $n_{2,b} V(n_{1,d}) (n_2 N_{2,l} N_{2,wd})^{-1}$ in (8) is almost always negligible. The estimates $\hat{N}_{2,d}$ and $\hat{N}_{2,l}$ of the sizes of the domains $U_{2,d}$ and $U_{2,l}$ can be obtained from a sample from the original or current survey population. If more than one sample is drawn, each can provide an unbiased estimate of $N_{2,d}$ (or $N_{2,l}$), all of which can be combined. The minimum variance combined estimator is the sum of the estimators weighted with the reciprocals of their variances. As the following argument shows, we do not expect the bias of (9) to be large:

$$\begin{aligned}
E(\hat{t}'_{y\pi}) &= E[\hat{t}_{y\pi} (1 + \hat{c})^{-1}] \approx E(\hat{t}_{y\pi}) (1 + c)^{-1} \\
&= t_y (1 + c) (1 + c)^{-1} = t_y.
\end{aligned}$$

Another strategy, here denoted by Strategy 3, is to feed back the information that certain units are dead, but to retain them on the frame and allow them to be sampled. The resulting estimator is unbiased, but the disadvantage of this strategy is that the precision will suffer as part of the sample is lost on ineligible units. The estimator of Strategy 3 is $\hat{t}''_{y\pi} = N_1 / n_2 \sum_r y_k$, where r is a sample from the original survey population U_1 .

3.2 A Simulation Study

A simulation study may shed some light on which of the Strategies 1–3 is to be preferred. Natural measures for comparing the strategies are bias and variance. In business surveys, estimates for subpopulations (industries) are often more interesting than the whole population. To simulate a subpopulation, a frame consisting of 1,000 units was created to form the original survey population. A gamma distributed value, Y1, was associated with each unit. We used the same gamma distribution as the one that generated Population 12 in Lee, Rancourt and Särndal (1994, page 236). The coefficient of variation (population standard deviation divided by the mean) was 0.57. Another study variable, Y2, was created by performing independent Bernoulli trials, one for each population unit, which obtained value 1 with probability equal to 0.5 and value 0 otherwise. Unlike in Lee *et al.*, some of the units were dead. Each unit was independently of other units classified as dead with a probability P_{dead} . All dead units were assigned zero values for both Y1 and Y2. A set of Y1 and Y2 were simulated for each of four values of P_{dead} : 0.03, 0.05, 0.2, and 0.5. These sets contained 29, 54, 201 and 494 dead units, respectively.

A PRN was attached to each unit and the units were laid out along a PRN line. The first sample, s_1 , was drawn by identifying the 500 units with the smallest PRNs. All dead units in s_1 were flagged as ‘dead by sample survey sources’. Hence, s_1 covered approximately the first half of the PRN line. The frame with the units flagged as dead by sample survey sources excluded made up the current survey population. The estimates of $N_{2,d}$ and $N_{2,l}$ used in Strategy 2 were based on s_1 . A second sample, denoted by $s_{2\text{current}}$, was drawn by taking 100 units to the right of a starting point, *start 2*, disregarding units dead by sample survey sources. Another sample of 100 units was selected from *start 2*, but units dead by sample survey sources were this time allowed to be included in this sample. Hence, this sample was drawn from U_1 , and we denote it by $s_{2\text{orig}}$. The sample $s_{2\text{current}}$ is pertinent to Strategies 1 and 2 while $s_{2\text{orig}}$ will be used for Strategy 3.

The procedure described in the preceding paragraph was repeated 1,000 times. That is, for each of the values of P_{dead} mentioned above and for each of three starting points of s_2 , to be defined, 1,000 sets of PRNs were generated and attached to the units. The frame was reordered for each new

set of PRNs, and three samples were drawn for each reordering (s_1 , $s_{2\text{current}}$, and $s_{2\text{orig}}$). Two values of *start* 2, 0.0 and 0.7, were chosen so as to make the proportion of $s_{2\text{current}}$ that fell in $s_{1,l}$ 100% and 0%, respectively. That is, $n_{2,a}/n_2$ was set to 100% and 0%. Further, to make $n_{2,a}/n_2$ on average 50% under each of the chosen P_{dead} , appropriate values of *start* 2 were derived. They are 0.448, 0.447, 0.438, and 0.4 for the P_{dead} values 0.03, 0.05, 0.2, and 0.5, respectively.

In summary, the population and samples sizes, the study variables Y1 and Y2, and which of the units that were dead were held fixed in our study. For twelve combinations of P_{dead} and $n_{2,a}/n_2$, the reordering of the units on the PRN line through the simulation of new PRNs made the following factors vary:

- which of the units that were included in s_1 , $s_{2\text{current}}$, and $s_{2\text{orig}}$;
- how many and which of the dead units that were dead by sample survey sources;
- which of the units that belonged to $s_{1,l}$ and $U_{2,wd}$.

Thus the quantities $s_{1,d}$, $N_{2,d}$ and N_2 vary in the simulations. It seems practical to let them do so rather than controlling them in an experiment with more factors than

P_{dead} and $n_{2,a}/n_2$. Hence the results are unconditional, in accordance with (8).

3.3 Results

Table 1 shows the empirical relative bias of Strategies 1 and 2, computed as the straight average of the 1,000 differences between the estimate and the parameter in terms of the percentage of the total obtained in the simulation. Strategy 3 is unbiased and is therefore not included in Table 1. The empirical bias of Strategy 3 that nevertheless appeared in the simulations reflects the simulation error; it was at most 0.5%. As seen in Table 1, Strategy 2 is virtually unbiased as well. Note that the simulated empirical bias under Strategy 1 is what (8) predicts (with allowance for simulation error). This bias is appreciable in nearly all cases and if the proportion of dead (or ineligible) units is high the bias can be very severe indeed. Figure 2 shows the conditional bias given $n_{1,d}$ for $P_{\text{dead}} = 0.50$ and $n_{2,a}/n_2 = 0\%$. Note that the bias given by (6) is locally well described by the regression line in the figure defined by the OLS fit of the bias conditional on $n_{1,d}$. For example, if $n_{1,d} = 220$, then both $N_{2,d}/N_{2,l}$ and $(n_1 - n_{1,d})/(N_1 - n_1)$ equal 0.56 and $B = -0.31$.

Table 1
Bias, % of Total of Y1. The First Entry in Each Cell is the Bias Under Strategy 1, the Second is the Bias Under Strategy 2.

P_{dead}	Average of n_a/n					
	0%		50%		100%	
0.03	-1.6	-0.1	0.4	0.4	1.5	0.0
0.05	-2.8	0.0	0.4	0.4	2.9	0.0
0.20	-10.2	-0.2	1.5	0.4	12.7	0.1
0.50	-24.6	0.2	12.5	0.3	49.0	0.2

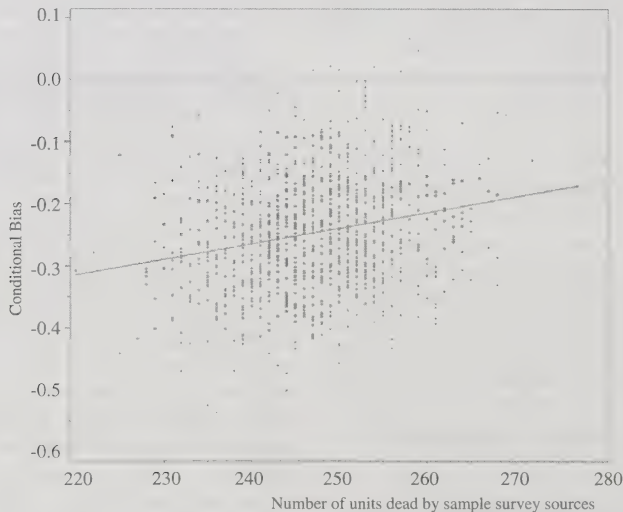


Figure 2. The simulated conditional bias plotted against the number of units dead by sample survey sources, $n_{1,d}$, for $P_{\text{dead}} = 0.50$ and $n_{2,a}/n_2 = 0\%$. An OLS regression line shows the local trend of the conditional bias as a function of $n_{1,d}$.

To assess the bias it helps to look at the coverage probabilities. Table 2 shows the empirical coverage probabilities, based on symmetric 'confidence intervals' with a width of two times the simulated empirical standard deviation of each side of the point estimate. While Strategy 2 gives in all cells coverage probabilities close to the targeted 95%, Strategy 1 achieves that in general only for the population with 3% dead units. The coverage probability under Strategy 1 tends also to be acceptable for populations with a larger proportion of dead units, if half of the sample is taken from the part of the PRN line where dead units have been weeded out, and the other half from the part of the PRN line where the original proportion of dead units has been retained, as the negative bias from the first half of the sample tends to cancel out the positive bias from the second half.

The variance of the simulated estimates was computed. Tables 3 and 4 show the variance comparisons for Y1 and Y2, respectively, under Strategies 2 and 3 relative to that of Strategy 1. As expected, in all cases Strategy 1 gave a smaller variance than did Strategy 3. Strategy 2 performed well in most cases, but considering the extra complexity of this strategy, the feed back Strategy 1 seems preferable for populations with a small proportion of ineligible units, say 3% or less. However, if this proportion is larger than, say, 5%, the bias of Strategy 1 may cause poor coverage probabilities and misleading estimates. The variance of Strategy 2 is no worse than that of Strategy 3; in most cases Strategy 2 is superior. The non-monotone variance ratios in the bottom row of Table 3 is due to the estimation of $N_{2,d}$ and $N_{a,l}$ combined with the specific details of the simulation.

Table 2

The Coverage Probability in Percentage for Estimating Total of Y1. The First Entry in Each Cell is the Coverage Probability Under Strategy 1, the Second is the Coverage Probability Under Strategy 2.

P_{dead}	Average of n_a/n					
	0%		50%		100%	
0.03	94.6	94.3	94.6	94.8	94.3	95.1
0.05	93.3	95.2	94.4	93.9	90.8	95.0
0.20	65.9	94.5	93.8	94.8	46.1	94.6
0.50	21.2	95.1	78.4	94.7	0.0	94.8

Table 3

Variance Ratio of the Estimator of the Total of Y1. The First Entry in Each Cell is the Variance Under Strategy 2 Relative to that of Strategy 1, the Second is the Variance Under Strategy 3 Relative to Strategy 1.

P_{dead}	Average of n_a/n					
	0%		50%		100%	
0.03	1.04	1.04	1.00	1.06	0.98	1.08
0.05	1.08	1.08	0.98	1.14	0.95	1.15
0.20	1.28	1.28	0.85	1.27	0.83	1.46
0.50	1.85	1.85	0.52	1.34	0.58	2.24

Table 4

Variance Ratio of the Estimator of the Total of Y2. The First Entry in Each Cell is the Variance Under Strategy 2 Relative to that of Strategy 1, the Second is the Variance Under Strategy 3 Relative to Strategy 1.

P_{dead}	Average of n_a/n					
	0%		50%		100%	
0.03	1.03	1.03	1.00	1.03	0.97	1.03
0.05	1.06	1.06	0.99	1.04	0.95	1.06
0.20	1.25	1.25	0.92	1.15	0.80	1.19
0.50	1.80	1.81	0.65	1.40	0.50	1.36

4. DISCUSSION

This paper gives conditional and unconditional expressions for the feed back bias when the total is estimated with the common expansion estimator. We have shown that the feed back bias can be large. With as little as 5% ineligible units on the frame, feeding back information of these from sample surveys can result in about 2–3% bias. However, a small-scale simulation study indicates that if the proportion of ineligible units is 3% or less, the feed back strategy does not seem to create problems in terms of bias and variance.

We have also derived a virtually unbiased estimator. The simulation study shows that this estimator compares favourably in terms of variance with the alternative strategy of retaining ineligible units on the frame and letting them be included in further samples. This estimator relies on the availability of consistent estimates of the number of eligible and ineligible units in the population. These estimates may be obtained from an earlier sample through the unbiased strategy of letting units that have been found dead be included in the sample.

In order to facilitate the theoretical development, we have made simplifying assumptions. The most important of these is the assumption that *all* dead units have been found in earlier sample surveys and have been fed back to the frame. We have envisaged a frame with one 'white' area, where all ineligibles have been flagged as such, and one 'black' area, where no ineligibles have been touched. In practice, this is not likely to happen. If the frame is moderately large and used for many continuing surveys, some of which may feed back to varying intensity, the frame will turn 'grey' rather than 'black and white'. The feed back bias will then be less severe than in the 'black and white' situation. It has not, however, been in the scope of this paper to quantify the bias for a 'realistically grey' frame. In this sense, what has been examined in this paper is a worst case scenario.

ACKNOWLEDGEMENTS

The authors thank Mark Pont for very useful initial discussions of this topic. They are also most grateful to an associate editor and two referees for very valuable comments. Both authors' research was partially supported by the UK Office for National Statistics and Wang's research was also supported by the U.S. National Cancer Institute

(CA 57030). Hedlin was employed by University of Southampton when he took part in this work.

REFERENCES

- COLLEDGE, M.J. (1989). Coverage and classification maintenance issues in economic surveys. In *Panel Surveys*, (Eds., D. Kasprzyk, G.J. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc., 80-107.
- COLLEDGE, M.J. (1995). Frames and business registers: an overview. In *Business Survey Methods*, (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 21-47.
- DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, (Eds., N.L. Johnson and H. Smith). New York: John Wiley & Sons, Inc., 629-651.
- ERNST, L.R., VALLIANT, R. and CASADY, R.J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics*, 16, 211-228.
- HIDIROGLOU, M.A., and LANIEL, N. (2001). Sampling and estimation issues for annual and sub-annual Canadian business surveys. *International Statistical Review*, 69, 487-504.
- HIDIROGLOU, M.A., and SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.
- LEE, H., RANCOURT, E. and SÄRNDAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- LAVALLÉE, P. (1996). Frame update problems with panel surveys. *Proceedings of Statistical Days '96*, Statistical Society of Slovenia, 252-261.
- OHLSSON, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods*, (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 153-169.
- SCHIOPU-KRATINA, I., and SRINATH, K.P. (1991). Sample rotation and estimation in the survey of employment, payrolls and hours. *Survey Methodology*, 17, 79-90.
- SRINATH, K.P. (1987). Methodological problems in designing continuous business surveys: some Canadian experiences. *Journal of Official Statistics*, 3, 283-288.
- SRINATH, K.P. and CARPENTER, R.M. (1995). Sampling methods for repeated business surveys. In *Business Survey Methods*, (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge and P. Kott). New York: John Wiley & Sons, Inc., 171-183.
- THOMPSON, M.E. (1997). *Theory of Sample Surveys*. London: Chapman & Hall.

Application of Quality Control in ICR Data Capture: 2001 Canadian Census of Agriculture

WALTER MUDRYK and HANSHENG XIE¹

ABSTRACT

Intelligent Character Recognition (ICR) has been widely used as a new technology in data capture processing. It was used for the first time at Statistics Canada to process the 2001 Canadian Census of Agriculture. This involved many new challenges, both operational and methodological. This paper presents an overview of the methodological tools used to put in place an efficient ICR system. Since the potential for high levels of error existed at various stages of the operation, Quality Assurance (QA) and Quality Control (QC) methods and procedures were built into this operation to ensure a high degree of accuracy in the captured data. This paper describes these QA / QC methods along with their results and shows how quality improvements were achieved in the ICR Data Capture operation. This paper also identifies the positive impacts of these procedures on this operation.

KEY WORDS: Data Capture; Intelligent Character Recognition (ICR); Quality control; Quality improvement; Statistical process control.

1. INTRODUCTION

The data capture of the 2001 Canadian Census of Agriculture was conducted between July and November 2001, using relatively new technology called Intelligent Character Recognition (ICR). This approach to data capture combines Automated Machine Capture which uses optical character, mark and image recognition, with Manual Capture by operators who 'key from image' using a heads-up data capture technique. The heads-up data capture technique is applied only to fields that can not be recognized by the optical system with a sufficiently high degree of confidence (that is pre-specified).

The ICR system offered many benefits to the data capture operation, in terms of resource savings and productivity gains. At the same time, accuracy became an extremely important consideration for processing a large number of documents since the potential for unacceptable levels of error existed at various stages of the process. In the literature, the quality of ICR applications has been studied by a few authors; see, *e.g.*, Kalpic (1994) and Pasley (2000), among others. Kalpic discussed the coding algorithm and the results for the 1991 Census Coding Operation in Croatia and Bosnia-Herzegovina, using intelligent optical readers. Pasley pointed out that the quality of a scanned image usually depends on the quality of the source document, the precision of the scanner, the skill of the scanner operator and the resolution at which the document was scanned. With quality improvement in mind, QA and QC procedures were built into the data capture operation for the 2001 Canadian Census of Agriculture to ensure a high degree of accuracy in this operation.

Quality Control activities for the ICR Data Capture Operation were focused in three main stages of processing, namely: document preparation, scanning calibration, and data capture of the questionnaires. This was done since each of these stages was dependent on one another and each had the potential to contribute significant errors down the line. Therefore, each component should ideally have its own control system.

It is the purpose of this paper to describe the QA/QC methodology and procedures associated with each of the main stages of the ICR Data Capture Operation, summarise the results obtained from their application and show how ongoing quality improvements were achieved in the ICR Data Capture operation.

2. QUALITY PROGRAM OVERVIEW

To better understand the rationale behind the QA/QC procedures, it is worthwhile to give an overview of their objectives and methodologies.

2.1 Objectives

The overall quality objective for this project was to measure, control and improve the quality of the entire ICR Data Capture Operation on a continuous basis. This would be achieved by implementing a series of QA/QC procedures at each critical stage of the operation. The specific objectives for each stage were as follows:

- a) Document Preparation: to ensure that only highly readable documents would reach the scanning stage.

¹ Walter Mudryk and Hansheng Xie, Business Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6.

- b) Scanning Calibration: to ensure optimal machine set-up and calibration prior to the start of production.
- c) Quick Capture (Machine Capture) and Quick Key (Manual Capture): to ensure a high level of quality of data capture during production.

2.2 QA / QC Methodologies

Each major stage of processing was operationally unique and therefore, had different quality requirements. As a result, QA procedures were applied to the Document Preparation operation, and QC procedures to the Scanning Calibration, Quick Capture and Quick Key operations. A flowchart is given in the Appendix, which shows the various stages of the ICR Data Capture Operation and exactly where these procedures were applied.

2.2.1 Document Preparation

The document preparation operation was essentially divided into five sub-processes, specifically: sorting, transcription, batching, cutting and storage. This operation was responsible for preparing the questionnaires and associated batches for scanning by the ICR equipment and was performed manually by clerical staff. It included activities such as separating the contents of the received envelopes by document type (*Sorting*), re-transcribing damaged or illegible questionnaires (*Transcription*), grouping questionnaires into batches for registration (*Batching*), cutting the spine of each booklet questionnaire with an electric cutter (*Cutting*) and filing questionnaires in the archive (*Storage*). One of the most important aspects of this operation was the identification and isolation of problematic questionnaires so that they would not advance undetected to the scanning and data capture stages. These problematic questionnaires were labeled as 'outlier' questionnaires since they had problems such as questionnaires being X'ed out or written over fields, extraneous markings, illegible entries, torn, crumpled or taped documents, etc.

The potential for error in this operation could lead to some problems being experienced at the scanning stage. It was felt that QA procedures would be appropriate to ensure quality at this stage since many of the clerical functions were also subject to various automated system cross-checks. The system cross checks ensured that the documents had a valid ID, correct number of pages, and that the pages, once cut, were aligned and in sequential order. The QA procedures consisted of a series of on-going random spot checks for each of the five sub-processes. The results of each spot check were recorded on a control form and summarized for the supervisor to identify if the work was being done correctly. Feedback would then be given to the individual clerk or group on a regular basis, and corrective actions would be taken when necessary. For example, if the

work was not being performed well, some re-training would take place and/or an increase in the frequency of spot-checks was done until favorable results were obtained. If extensive problems were identified, the supervisor could also decide on the amount of re-work required, based on the seriousness of the problem observed.

For the sorting, batching, cutting and storage operations, the quality measure selected was '*percent of questionnaires in error*' (i.e., in keeping with the assumptions required for a simple sampling unit). For the transcription operation, the probability of multiple independent errors occurring within a questionnaire was extremely high and therefore the quality measure selected was '*Defects per Hundred Units, DPHU*' (i.e., in keeping with the assumptions required for a complex sampling unit).

2.2.2 Scanning Calibration Check

Experience has shown that if the scanning equipment is not properly configured, the potential for generating poor quality images increases substantially. It is therefore imperative that the scanning equipment be optimally set prior to production and well maintained throughout the scanning operation. To ensure this, a QC procedure called the Scanning Calibration Check was developed to review the machine settings and calibration on an ongoing basis.

Since the equipment settings of the scanning system would tend not to fluctuate too greatly, it was felt that Statistical Process Control (SPC) methods would be appropriate for controlling this portion of the operation. This would essentially be an ongoing spot check of the calibration settings performed on a daily basis prior to the start of production. The calibration check consisted of re-scanning a test batch and comparing the results with the corresponding pre-benchmarked results for the same batch. The differences between the actual and expected results would be compared and error rates computed. These error rates were then plotted on SPC control charts to determine if the process was operating at an acceptable level. If this test batch failed, the scanning process would not be allowed to start production until the machine was re-calibrated and subsequently re-tested successfully.

In the Scanning operation, machine recognition could substitute wrong values when poor quality images are produced. Poor images could be the result of many factors such as dirty read heads, smeared optical windows, misalignment, mis-registration of fields, poor contrast / brightness levels, paper feed problems, etc. Since a specific quality standard was established for each field type, a separate *p* control chart was used to evaluate the substitution error rate for each type (specifically, alpha, alphanumeric, numeric, tick boxes and bar codes). The acceptable quality standard for each field type was previously established on a

field type basis by the client area so therefore, the quality measure used was 'percent of fields in error', i.e., the substitution error rate by field type for each scanner.

Based on SPC control chart theory, a decision for each scanning calibration test was made as follows:

- If each of the sample error rates for the five field types was respectively lower than their corresponding upper control limit (UCL), it was concluded that the scanning system was functioning properly and was ready for scanning production.
- Otherwise, it was concluded that a problem existed with the scanning equipment, and corrective action must be taken before the start of regular production.

The test batches were constructed with minimum sample size requirements in mind for each field type, such that the producer's confidence level would be at least 95%. This was then used as a guide in selecting the actual questionnaires for each of the test batches. The *minimum size* was required for each field type in order to achieve the high efficiency of decisions in the scanning calibration test, while the *Producer's Confidence Level* referred to the likelihood that the scanning system would pass the test for that field type when the system was functioning at the acceptable target level. The Upper Control Limit for each field type was computed assuming a $+2\sigma$ variability. This limit is lower than the customary $+3\sigma$ Upper Control Limits since the scanning calibration check was designed to be more sensitive in detecting smaller shifts at start-up than during normal production.

2.2.3 Quick Capture and Quick Key

Once the questionnaires had been scanned, the system would produce a digital image of each field along with an interpretation of its value and an associated confidence level for its recognition. The actual data capture then consisted of two processes: Quick Capture and Quick Key. Quick Capture was the automatic recognition by the system of all field images whose confidence levels were above a pre-specified threshold value. Quick Key consisted of the heads-up manual capture (by keyers working on terminals) of field images whose confidence levels were below the pre-set threshold value.

Since under ideal circumstances, these two processes were expected to be relatively stable, the QC Procedures were again based on SPC principles and were developed to measure and monitor the quality of each of the processes. This QC approach consisted of a small sample check from the output of a sample of batches taken systematically over time and computing the error rates for each sample. These error rates would then be compared to rejection levels that were calculated by the system based on the expected quality standard and the sample size for that observation. A

decision was then made as to the acceptability of each of these sample measurements relative to the expected quality standard for that process.

In the case of the Quick Capture operation, the machine may interpret a different value from the actual value for that field, and therefore, substitution rates were used to evaluate this process. These substitution errors are particularly serious since, if left unchecked, they may affect the recognition rate for many fields for a long period of time. In the case of the Quick Key operation, operators may make keying errors for many reasons such as lack of skill, poor training, fatigue, etc., and therefore, keying error rates were used to evaluate this manual process. For both of these processes, the quality measure was defined as 'percent of fields in error', across all field types combined.

Within the two capture operations, there were two distinct categories for processing the scanned documents: *Regular* questionnaires and *Outlier* questionnaires. QC procedures were put in place for each category. A separate sample was required for each process, one for Quick Capture and one for Quick Key. The system could distinguish between Quick Capture and Quick Key fields in each sample questionnaire and maintain separate counts of these fields that had been captured under each process. These field counts eventually became the sample size for each sample. Each sample was then compared to its own *threshold rejection rate*, which was a function of the number of fields observed (i.e., the effective *sample size*) and the expected quality standard or target for that process. A decision would then be made to accept or reject the sample. The threshold rejection rate was equivalent to the standard Upper Control Limit (UCL) that would be calculated on a standard p control chart. If the sample error rate exceeded this level, the process was rejected and the QC Reviewer proceeded to investigate and implement corrective actions as appropriate; otherwise the process was accepted.

The sampling was done on an individual scanner basis for Quick Capture and an individual operator basis for Quick Key. Some operators required more questionnaires to be sampled from time to time, and others less, based on their actual performance. Since the actual observations were based on samples, a customary $+3\sigma$ variability was permitted above the expected quality standard (i.e., the centerline of a p control chart) for each process. The batch decisions for these sample observations were made by the system during QC verification and these results were then plotted on a p control chart for each scanner and operator, after the fact and updated weekly.

For a detailed description of these QA/QC procedures and their rationale, please refer to Mudryk, Bougie and Xie (2001).

3. QUALITY IMPROVEMENTS

Two essential elements were included in the quality improvement strategy for the ICR Data Capture Operation. These consisted of feedback of QA/QC results and the implementation of corrective and preventive actions when required. These two elements enabled various staff to play an active role in improving the quality of each process through the additional insight into the problems that were identified and through the subsequent corrective or preventive actions that were taken.

Using QC data analysis as the base, all processes were examined to determine if they were operating efficiently. QC meetings were held with operations staff on a weekly basis to review the ongoing progress of the entire operation. Problems that had impacted any of the processes were addressed and recommendations made to treat their root causes and prevent their re-occurrence. The involvement of operational staff in resolving these problems played an important part in facilitating quality improvements on a continuous basis. The following examples illustrate some of the more significant corrective actions that were taken during the operation that led to quality improvements at various stages.

Example 1: Filtering Process for Detecting Outlier Documents

During the first few weeks of production, it was noticed that some documents were causing a high concentration of errors from things like large X's across a page, 0's and dashes in various fields, etc. These documents were causing high error rates for both operations but especially for the Quick Capture process. Since these documents were very different from the majority of the regular documents, a procedure was introduced to sort these documents for special treatment and processing after the fact. Some documents in fact had to be re-transcribed at this stage prior to processing them by ICR.

Example 2: Adjusting System Settings for Scanning & Recognition

The highlights of the QC weekly summaries indicated that both scanners made errors frequently on Pages 3 and 14 of the questionnaires during the first few weeks of processing. An investigation was conducted and it was found that there was a template reading problem on Page 3 and the pre-set recognition threshold level for the numeric fields on Page 14 were set too low. After the system settings on both scanners were adjusted, the system showed substantial improvements in the scanning of these two pages.

Example 3: Retraining Operators with High Error Rates

During the keying operation, the QC results showed that certain keyers were experiencing above average difficulties with the 'key from image' process and that their error rates

remained high for several weeks. Focusing on continuous improvement, these keyers were offered retraining on an ongoing basis. As a result, many keyers made significant improvements (week by week) in their keying performance.

4. QC EVALUATION AND ANALYSIS

Throughout the operation, many QC reports, charts and estimates, were produced to provide information about the incoming and outgoing quality levels and to evaluate the output of each production process. These reports were used to analyse the quality of each process by week and across weeks.

4.1 Document Preparation

For each of the five sub-processes of the document preparation, individual QA procedures were applied at different frequencies and both corrective and preventive actions were taken on an on-going basis as dictated by the results. The information collected and the feedback that was provided as a result of these QA procedures helped significantly in improving the scanning, imaging, recognition and capture of the questionnaires. In the first few weeks of production, it was discovered from the QC results that problematic documents (*i.e.*, outliers) were causing most of the substitution errors (*i.e.*, machine errors) in the *Quick Capture* process. From that point on, a new procedure was introduced into the *Sorting* process of the Document Preparation operation to separate these documents for special treatment from the regular documents (*i.e.*, labeled them for subsequent 100% verification). In general, better quality documents reached the scanning stations while poorer documents were either re-transcribed or processed separately with the addition of post processes such as 100% verification.

4.2 Scanning Calibration Check

In an effort to ensure optimal scanner settings and calibration, a *Scanning Calibration Check* was initially conducted twice a day, and subsequently once a day, prior to production processing. Many test batches were scanned during the operation with a relatively high rejection rate encountered by each scanner. On average, approximately 2-3 tests per day (with corresponding re-calibrations) were required for optimising the set-up of each of the two scanners. This demonstrates the need for re-calibration between processing periods. It should be noted that some rejections occurred due to problems identified with the test batches which were fixed later on. This is definitely an area where some procedural improvement is required in the future.

Both scanners exhibited reasonably high variability during this test. The high number of tests required, high rate of rejection and high variability across processing periods for many of the field types demonstrate the need to calibrate the scanning equipment properly prior to production. Otherwise, the scanners could be inadvertently set up to produce poor images right from the start, which would make good quality capture very difficult. Once a test batch failed, problems were usually identified and subsequent maintenance and corrective actions taken. This included actions such as: re-configuring the scanning equipment, replacing old light bulbs, fixing software problems, cleaning dirty read heads, *etc.* Using this test, the scanners were able to be calibrated and maintained at optimum levels of performance, between production runs.

4.3 Quick Capture and Quick Key

For the *Quick Capture* process, over the entire 18 weeks of processing the *Regular questionnaires*, the overall weekly substitution error rates decreased steadily from 4.3% to 0.8%, resulting in a grand overall substitution error rate of 2.0% (across all field types) for both scanners. The substitution error rates measured during production were maintained very near the Target levels that were established for each field type. These were as follows: Alpha (2.1% relative to a target of 2.0%); Alphanumeric (3.2% vs. 3.5%); Bar Code (0.0% vs. 0.2%); Numeric (2.8% vs. 2.0%) and Tick Boxes (0.8% vs. 0.4%). In comparison, processing the *outlier questionnaires* had a much higher substitution error rate and greater weekly variability than the corresponding *regular* questionnaires (*i.e.*, ranged from a high of 22.4% to a low of 1.3%). Although the substitution error rate did tend to reduce substantially over time, it did remain relatively high throughout the process and was measured at 7.0% overall, which was significantly higher than the rate for *regular* questionnaires (*i.e.*, 2.0%).

For the *Quick Key* process, the keying error rate for processing the *regular questionnaires* was relatively high

throughout the entire processing period (*i.e.*, mostly over 3%). This was partially due to the fact that this operation was a *heads-up* keying process and these keyers typically processed the most difficult cases. Over the entire 18 weeks however, the weekly keying error rates generally decreased from 5.6% to 1.6%, with an overall average of 3.4%. The keying was also subject to high levels of variability among operators, with individual error rates ranging 1.7% to 7.5%. It is interesting that keying the *outlier questionnaires* had a similar keying error rate to the corresponding *regular* process (*i.e.*, 3.4% vs. 3.7%) and ranged from a high of 5.7% to a low of 1.6%.

4.4 Estimates of Average Outgoing Quality

The primary purpose of the QA/QC procedures was to identify problems and to prevent them from occurring again. However, these procedures also had a corrective component in the sense that, errors that were discovered were always rectified. It is therefore possible to estimate the overall Average Outgoing Quality (AOQ) for the data capture component after the application of the QC procedures.

Estimates of AOQ were calculated for each of the two data capture processes. For a sampled outlier batch, all the questionnaires (*i.e.*, sampled and remainder) in that batch would be subjected to subsequent 100% verification, while for a regular batch, only the sampled questionnaires would be verified. This affects the calculation of AOQ since it can be assumed that the outgoing error rate for all verified questionnaires is 0.0%. The overall estimate for each component was based on the information obtained from both the regular and outlier documents, considering estimates of incoming quality and corrections made during verification. In the calculation, any documents reprocessed through either Quick Capture or Quick Key were included in the count.

Table 1 provides estimates of the AOQ for the *Quick Capture* and *Quick Key* processes.

Table 1
Estimates of AOQ for ICR Data Capture

Process	No. Questionnaires in Population	No. Fields in Population	Estimated No. Fields Verified and Corrected	Incoming Error (%)	AOQ (%)
Quick Capture					
Regular	273,818	21,248,277	170,249	2.01	1.99
Outlier	12,702	1,044,358	1,044,358	6.99	0.00
Overall	286,520	22,292,635	1,214,607	2.95	1.90
Quick Key					
Regular	281,502	6,376,020	234,253	3.41	3.28
Outlier	25,788	686,734	686,734	3.67	0.00
Overall	307,290	7,062,754	920,987	3.45	2.97
Combined					
Regular		27,624,297	404,502	2.82	2.29
Outlier		1,731,092	1,731,092	5.09	0.00
Overall		29,355,389	2,135,594	3.24	2.16

It can be seen that the overall AOQ for the *Quick Capture* process was estimated at 1.90% and for the *Quick Key* process at 2.97%. This was down considerably from their corresponding estimates of incoming quality of 2.95% and 3.45% respectively. The overall AOQ for both processes was estimated at 2.16% (relative to an overall incoming error quality of 3.24%). It should be noted that the AOQ for *outlier* documents was assumed to be 0% since all *outlier* documents were subsequently 100% verified.

4.5 QC Summary

The above results clearly indicate the need for the QA/QC procedures at the different stages of processing. It also shows how they collectively contributed to controlling the outgoing quality and generating quality improvements into all phases of the ICR data capture operation.

The QC results clearly showed that the *outlier* documents had a greater negative impact on the *Quick Capture* process (i.e., 7.0% substitution error rate) than the *Quick Key* process (i.e., 3.7% keying error rate). This indicates that the filtering process for special treatment of *outlier* documents was an important step to take. The QC results also showed that if the documents were in good shape for scanning and the machines were well calibrated, the *automated* system was capable of capturing the data faster and with better quality than the manual *key from image* process. This is quite an important observation, since there are obvious savings implied with a corresponding improvement in data capture quality (i.e., 2.0% vs. 3.4%). To the defence of the keyers, however, they did process the more difficult cases, thus partially explaining their higher error rates. Overall, it was estimated that about 77% of the fields were captured through the *Quick Capture* process and 23% were captured through the *Quick Key* process.

It should also be noted that the regular feedback of the QC information collected from the various stages of the ICR process was essential in identifying the root causes of many problems and in helping to resolve them. This provided the opportunity for many quality improvements to be generated into the various stages, on an on-going basis.

For a detailed description of these QA/QC results, please refer to Mudryk and Xie (2002).

5. CONCLUSIONS

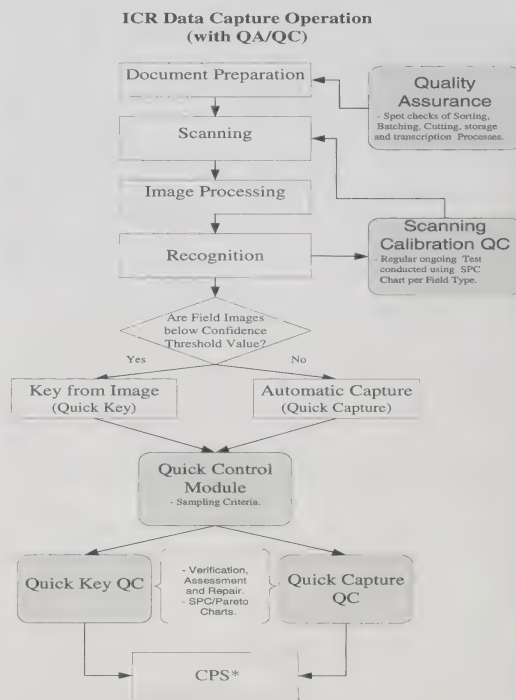
It is clear from the results obtained in this analysis, that the QA/QC procedures were extremely valuable and had a very positive impact on the entire operation. The QA procedures that were applied in the Document Preparation process were effective in preventing many poor documents from reaching the scanning stations and those that did were

then labeled for special treatment and subsequent 100% verification.

The QC procedures were then able to optimize the machine set-up by applying the Scanning Calibration Check prior to production. Furthermore during production, QC samples were also able to identify problems with the automatic recognition and key from image processes, so that they could be improved as required.

In all cases, early warning signals were obtained from objective measurements at each stage of processing, and corrective and preventive actions were implemented as needed. Extensive feedback was provided to all stages of the ICR process on an ongoing basis from which continuous quality improvements were generated.

APPENDIX



* CPS = Central Processing System.

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, to an Associate Editor and to an Assistant Editor for their detailed and constructive comments. They also thank Bob Bougie for many helpful comments.

REFERENCES

- KALPIC, D. (1994). Miscellanea, Automated Coding of Census Data. *Journal of Official Statistics*, 10, 4, 449-463.
- MUDRYK, W., BOUGIE, B. and XIE, H. (2001). Quality Control of ICR Data Capture: 2001 Canadian Census of Agriculture. *International Conference on Quality in Official Statistics in Stockholm*, Sweden.
- MUDRYK, W., and XIE, H. (2002). Quality Control Application in ICR Data Capture for the 2001 Canadian Census of Agriculture. *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2424-2429.
- PASLEY, B. (2000). Web Exclusive: The Good and Bad of Scanned Images. Posted on the POB (Point of Beginning) website.

Design Effects for the Weighted Mean and Total Estimators Under Complex Survey Sampling

INHO PARK and HYUNSHIK LEE¹

ABSTRACT

We revisit the relationship between the design effects for the weighted total estimator and the weighted mean estimator under complex survey sampling. Examples are provided under various cases. Furthermore, some of the misconceptions surrounding design effects will be clarified with examples.

KEY WORDS: Simple random sample; pps sampling; Multistage sampling; Self-weighting; Poststratification; Intraclass correlation coefficient.

1. INTRODUCTION

The design effect is widely used in survey sampling for developing a sampling design and for reporting the effect of the sampling design in estimation and analysis. It is defined as the ratio of the variance of an estimator under a complex sampling design to that of the estimator under simple random sampling with the same sample size. An estimated design effect is routinely produced by computer software packages for complex surveys such as WesVar and SUDAAN. It was originally intended and defined for the weighted (ratio) estimator of the population mean (Kish 1995). However, a common practice has been to apply this concept for other statistics such as the weighted total estimator often with success but at times with confusion and misunderstanding. The latter situation occurs particularly when simple but useful results derived under a relatively simple sampling design are applied to more complex problems. In this paper, we examine the relationship between the design effects for the weighted total estimator and the weighted mean estimator under various complex survey sampling designs. In section 2, we briefly review the definition of the design effect and its practical usage while discussing some of the misconceptions surrounding design effects for the weighted total and mean estimators. Subsequently, in section 3, we analyze the difference between the design effect for the weighted total estimator and that for the weighted mean estimator under a two-stage sampling design followed by a discussion regarding the design effects under various two-stage sampling designs and some more general cases in section 4. We try to clarify some of the misconceptions with these examples. Finally, we summarize our discussion in section 5.

2. A BRIEF REVIEW ON DEFINITION AND USE OF DESIGN EFFECT IN PRACTICE

A precursor of the design effect that has been popularized by Kish (1965) was used by Cornfield (1951). He defined the efficiency of a complex sampling design for estimating a population proportion as the ratio of the variance of the proportion estimator under simple random sampling with replacement (srswr) to the corresponding variance under a simple random cluster sampling design with the same sample size. The inverse of the ratio defined by Cornfield (1951) was also used by others. For example, Hansen, Hurwitz and Madow (1953, Vol. I, pages 259 – 270) discussed the increase of the relative variance of a ratio estimator due to the clustering effect of cluster sampling over simple random sampling without replacement (srswor). The name, design effect, or Deff in short, however, was coined and defined formally by Kish (1965, section 8.2, page 258) as “the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements” (for more history, see also Kish 1995, page 73 and references cited therein).

Suppose that we are interested in estimating the population mean (\bar{Y}) of a variable y from a sample of size m drawn by a complex sampling design denoted by p from a population of size M . Kish's Deff for an estimate (\bar{y}_p) is given by

$$\text{Deff} = \frac{V_p(\bar{y}_p)}{(1-f)S_y^2/m} \quad (2.1)$$

where V_p denotes variance with respect to p , $f = m/M$ is the overall sampling fraction, and $S_y^2 = (M-1)^{-1} \sum_{k=1}^M (y_k - \bar{Y})^2$ is the population element variance of the

¹ Inho Park and Hyunshik Lee, Westat, Inc. 1650 Research Blvd., Rockville, MD 20850, U.S.A. E-mail: InhoPark@westat.com.

y-variable. Although the design effect was originally intended and defined for an estimator of the population mean (Kish 1995), it can be defined for any meaningful statistic computed from a sample selected by a complex sampling design.

The Deff is a population quantity that depends on the sampling design and refers to a particular statistic estimating a particular population parameter of interest. Different estimators can estimate the same parameter and their design effects are different even under the same design. Therefore, the design effect includes not only the efficiency of the design but also the efficiency of the estimator. Särndal, Swensson, and Wretman (1992, page 54) made this point clear by defining it as a function of the design (p) and the estimator ($\hat{\theta}$) for the population parameter ($\theta = \theta(y)$). Thus, we may write it as

$$\text{Deff}_p(\hat{\theta}) = \frac{V_p(\hat{\theta})}{V_{\text{srswor}}(\hat{\theta})}$$

where $\hat{\theta}'$ is the usual form of an estimator for θ under srswor, which is normally different from $\hat{\theta}$. For example, to estimate the population mean, one may use the weighted (ratio) mean $\hat{\theta} = \sum_s w_k y_k / \sum_s w_k$ with sampling weights w_k but $\hat{\theta}'$ would be the simple sample mean $\sum_s y_k / m$, where the summation is over the sample s . We will see the effect of particular estimators $\hat{\theta}$ on the design effect in the later sections.

Kish (1995) later advocated using a somewhat different definition, which is called Deft and uses the srswr variance in the denominator on the ground that without-replacement sampling is a part of the design and should be captured in the definition. He also reasoned that Deft is easier to use for making inferences and that it is better to define the design effect without the finite population correction factor $(1-f)$ because the factor is difficult to compute in some situations. The new definition is given by

$$\text{Deft}_p(\hat{\theta}) = \sqrt{\frac{V_p(\hat{\theta})}{V_{\text{srswr}}(\hat{\theta})}}$$

or $\text{Deft}_p^2(\hat{\theta}) = V_p(\hat{\theta}) / V_{\text{srswr}}(\hat{\theta})$. Survey data software such as WesVar and SUDAAN produce Deft^2 instead of Deff. We will use this definition in this paper.

When the population parameter is the total (Y), the unbiased estimator is the weighted sample total, namely, $\hat{Y} = \sum_s w_k y_k$. When the population mean is the parameter of interest, it is usually estimated by the weighted mean, that is, $\hat{\bar{Y}} = \sum_s w_k y_k / \sum_s w_k$. It is a special case of the ratio estimator, $\sum_s w_k y_k / \sum_s w_k x_k$, where $x_k \equiv 1$ for all $k \in s$.

One common misconception about the design effects for \hat{Y} and $\hat{\bar{Y}}$ is that they are similar in values. However, it has been observed that the design effect for \hat{Y} , $\text{Deft}_p^2(\hat{Y})$,

tends to be much larger than that for $\hat{\bar{Y}}$, $\text{Deft}_p^2(\hat{\bar{Y}})$. This was also noted in, for example, Kish (1987) and Barron and Finch (1978). Some explanation can be found in Hansen *et al.* (1953, Vol. I, pages 336–340) who showed that the difference arises from the relative variance of the cluster sizes. More recently Särndal *et al.* (1992, pages 315–318) showed that contrary to the case of $\hat{\bar{Y}}$, the design effect for \hat{Y} depends on the (relative) variation of the y-variable. In fact, even the design effect for $\hat{\bar{Y}}$ may depend on the (relative) variation of the y-variable, which we will discuss in section 4. This dependence contradicts what the design effect is intended to measure as Kish (1995) explicitly described:

“Deft are used to express the effects of sample design beyond the elemental variability (S_y^2/m), removing both the units of measurement and sample size as nuisance parameters. With the removal of S_y , the units, and the sample size m , the design effects on the sampling errors are made generalizable (transferable) to other statistics and to other variables, within the same survey, and even to other surveys.”

His statement may be loosely true for the weighted mean $\hat{\bar{Y}}$ as expressed in the frequently used sample approximate formula for $\text{Deft}_p^2(p, \hat{\bar{Y}})$ given by Kish (1987):

$$\text{Deft}_p^2(\hat{\bar{Y}}) = \{1 + \rho(\bar{m} - 1)\} \{1 + \text{cv}_w^2\} \quad (2.2)$$

where the sample design p contains complex features such as unequal weighting and cluster sampling, $\rho = \rho_p(y)$ is the intraclass correlation coefficient (often called within cluster homogeneity measure), \bar{m} is the average cluster sample size, and cv_w^2 is the sample relative variance of the weights. Strictly speaking, this formula is not independent of the y-variable because ρ is dependent on the y-variable. Also, the design effect may not be free of the unit of measurement unless $V_p(\hat{\bar{Y}})$ is expressed in a factorial form of S_y^2/m . See Park and Lee (2002). This formula (2.2) is valid only when there is no correlation between the sampling weights and the survey variable y . However, if the correlation is present, the formula may need to be modified as studied by Spencer (2000) and Park and Lee (2001). In the following section, we elaborate this aspect in detail for two-stage sampling and we will also examine this point further in section 4.1.

3. DECOMPOSITION OF THE DESIGN EFFECT UNDER TWO-STAGE SAMPLING

We consider a sampling design conducted in two stages. Suppose that a population $U = \{k: k = 1, \dots, M\}$ with M elements is grouped into N clusters of size M_i such that

$M = \sum_{i=1}^N M_i$. The first stage sample $s_a = \{i: i=1, \dots, n\}$ of n clusters (primary sampling units, or PSUs in abbreviation) is selected with replacement from N clusters with probabilities p_i , where $\sum_{i=1}^N p_i = 1$. Let $p_a = \Pr(s_a)$ denote the first stage sampling design. The second stage sample $s_{bi} = \{j: j=1, \dots, m_i\}$ of m_i elements (secondary sampling units or SSUs in abbreviation) is then selected independently from each PSU i selected at the first stage according to some arbitrary sampling design, say $p_{bi} = \Pr(s_{bi} | s_a)$ where $i \in s_a$. Denote the total sample of elements and the overall sampling design by $s = \bigcup_{i \in s_a} s_{bi}$ and $p = \Pr(s)$, respectively. Associated with the j^{th} element in the i^{th} cluster is a survey characteristic y_{ij} , $j=1, \dots, M_i$, $i=1, \dots, N$. For a given $i \in s_a$, let $w_{j|i}$ be the second stage sampling weights such that an estimator of the form $\hat{Y}_i = \sum_{j=1}^{m_i} w_{j|i} y_{ij}$ is unbiased for the cluster total $Y_i = \sum_{j=1}^{M_i} y_{ij}$, that is, $E_b(\hat{Y}_i) = Y_i$, where E_b represents the expectation with respect to the second stage sampling. Let $w_i = 1/(np_i)$ be the first stage sampling weights and let $Y = \sum_{i=1}^N Y_i$ be the population total. It is easy to show that $E_a(Y_i / p_i) = Y$. Assuming that Y_i are known for $i \in s_a$, $\sum_{i=1}^n Y_i$ is the average of n unbiased estimators of Y so that $E_a(\sum_{i=1}^n w_i Y_i) = Y$, where E_a denotes the expectation with respect to the first stage sampling design. Note that both stages are sampling with replacement. Accordingly, it is possible that the same sampling unit (either cluster or element) is selected more than once but they are treated differently. Define the overall sampling weights by $w_{ij} = w_i w_{j|i}$. Clearly, $\hat{Y} = \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij} y_{ij}$ is unbiased for Y , that is, $E_p(\hat{Y}) = E_a E_b(\hat{Y}) = E_a(\sum_{i=1}^n w_i Y_i) = Y$, where E_p represents the expectation with respect to p . The variance of \hat{Y} can be written as

$$\begin{aligned} V_p(\hat{Y}) &= V_a E_b(\hat{Y}) + E_a V_b(\hat{Y}) \\ &= \sum_{i=1}^N w_i (Y_i - p_i Y)^2 + \sum_{i=1}^N w_i V_b(\hat{Y}_i) \end{aligned} \quad (3.1)$$

where V_p, V_a and V_b represent variances defined with respect to the overall, the first stage, and the second stage sampling. See Särndal *et al.* (1992, pages 151–152).

A commonly used estimator for the population mean $\bar{Y} = Y/M$ is the weighted (ratio) estimator given by $\hat{\bar{Y}} = \hat{Y}/\hat{M}$ where $\hat{M} = \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij}$. Using Taylor linearization, as shown in Särndal *et al.* (1992, pages 176–178), $\hat{\bar{Y}}$ can be approximated as

$$\hat{\bar{Y}} \cong \bar{Y} + M^{-1} \hat{D} \quad (3.2)$$

where $\hat{D} = \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij} d_{ij}$ is an unbiased estimator of the population total $D = \sum_{i=1}^N \sum_{j=1}^{M_i} d_{ij}$ of $d_{ij} = y_{ij} - \bar{Y}$, which represents the deviation of y_{ij} from the population mean \bar{Y} . Note that $D = 0$. Denoting $D_i = \sum_{j=1}^{M_i} d_{ij} = Y_i - M_i \bar{Y}$ and $\hat{D}_i = \sum_{j=1}^{m_i} w_{j|i} d_{ij}$, we obtain the approximate variance of $\hat{\bar{Y}}$ from expression (3.2) as

$$AV_p(\hat{\bar{Y}}) = \frac{1}{M^2} \left[\sum_{i=1}^N w_i \left(Y_i - \frac{M_i}{M} Y \right)^2 + \sum_{i=1}^N w_i V_b(\hat{D}_i) \right]. \quad (3.3)$$

If a simple random sample of size $m = \sum_{i=1}^n m_i$ is selected with replacement from the population U , then a sample mean $\bar{y}_{\text{srsw}} = \sum_{k=1}^m y_k / m$ and its expansion

$$\hat{Y}_{\text{srsw}} = M \bar{y}_{\text{srsw}} = \frac{1}{f} \sum_{k=1}^m y_k \quad (3.4)$$

would serve as the estimators of the population mean \bar{Y} and total Y , respectively, under srswr, where $f = m/M$ is the overall sampling fraction. Their variances under this sampling design are given as $V_{\text{srswr}}(\hat{Y}_{\text{srsw}}) = M^2 V_{\text{srswr}}(\bar{y}_{\text{srsw}})$, where $V_{\text{srswr}}(\bar{y}_{\text{srsw}}) = m^{-1} S_y^2$ and $S_y^2 = (M-1)^{-1} \sum_{k=1}^M (y_k - \bar{Y})^2$. We note that m is the achieved sample size, which is a random quantity in general. From (3.1), (3.3), and above expressions with m replaced by its expected value m' with respect to the overall sampling design p , i.e., $m' = E_p(m)$, the design effects for \hat{Y} and $\hat{\bar{Y}}$ can be written as

$$\text{Deft}_p^2(\hat{Y}) = \frac{m'}{CV_y^2} \left\{ \sum_{i=1}^N w_i \left(\frac{Y_i}{Y} - p_i \right)^2 + \sum_{i=1}^N w_i V_b \left(\frac{\hat{Y}_i}{Y} \right) \right\} \quad (3.5)$$

and

$$\text{Deft}_p^2(\hat{\bar{Y}}) \cong \frac{m'}{CV_y^2} \left\{ \sum_{i=1}^N w_i \left(\frac{Y_i}{Y} - \frac{M_i}{M} \right)^2 + \sum_{i=1}^N w_i V_b \left(\frac{\hat{D}_i}{Y} \right) \right\} \quad (3.6)$$

where $CV_y^2 = S_y^2 / \bar{Y}^2$ represents the population relative variance of the y -variable. From these expressions, the difference in design effects for \hat{Y} and $\hat{\bar{Y}}$ can be written as follows.

$$\text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\bar{Y}}) \cong \Delta_a + \Delta_b, \quad (3.7)$$

where

$$\Delta_a = \frac{m'}{CV_y^2} \left\{ \sum_{i=1}^N w_i \left[\left(\frac{Y_i}{Y} - p_i \right)^2 - \left(\frac{Y_i}{Y} - \frac{M_i}{M} \right)^2 \right] \right\}$$

and

$$\Delta_b = \frac{m'}{CV_y^2} \left\{ \sum_{i=1}^N w_i \left[V_b \left(\frac{\hat{Y}_i}{Y} \right) - V_b \left(\frac{\hat{D}_i}{Y} \right) \right] \right\}.$$

The two components Δ_a and Δ_b in expression (3.7) reflect the differences arising from the respective sources of variation from the first and second stages of sampling. Of course, the second component disappears if all the elements in selected clusters are observed since it becomes a single-stage design or if a simple random sample is selected in the second stage. This is because both variances $V_b(\hat{Y}_i)$ and $V_b(\hat{D}_i)$ are equivalent under the aforementioned conditions, that is, 1) $V_b(\hat{Y}_i) = V_b(\hat{D}_i) = 0$ if $w_{j|i} = 1$ for all i and j , and

2) $V_b(\hat{Y}_i) = V_b(\hat{D}_i) \geq 0$ if $w_{j|i} = M_i/m_i$ for all i and j . In other words,

$$\Delta_b = 0 \quad \text{if} \quad w_{j|i} = c_i \quad \text{for all } i \text{ and } j, \quad (3.8)$$

where c_i are nonnegative constants and not necessarily equal for different clusters. Meanwhile, we can show that

$$\Delta_a = \begin{cases} 0 & \text{if } p_i \propto M_i, \\ A_p(y) & \text{if } Y_i \propto M_i, \\ -A_p(y) & \text{if } p_i \propto Y_i, \end{cases} \quad (3.9)$$

for all i , where $A_p(y) = (m'/CV_y^2) \sum_{i=1}^N w_i(p_i - M_i/M)^2$. Note that $A_p(y)$ is a nonnegative quantity and also that the conditions in expression (3.9) can be restated, respectively, as $p_i = M_i/M$, $\bar{Y}_i = \bar{Y}$, and $p_i = Y_i/Y$, where $\bar{Y}_i = Y_i/M_i$ for all $i = 1, \dots, N$. This result reveals the effect of cluster sampling on the precision of the two estimators. For example, if $p_i = M_i/M$, cluster sampling makes no difference in the precision of the two estimators. On the other hand, if $p_i = Y_i/Y$, \hat{Y} becomes more efficient than $\hat{\bar{Y}}$ in precision under cluster sampling, whereas the cluster sampling favors $\hat{\bar{Y}}$ over \hat{Y} in terms of precision if $\bar{Y}_i = \bar{Y}$ for all i .

Now, let us consider some examples of the conditions of (3.8) and (3.9).

Example 3.1 For one or two-stage cluster design with pps cluster sampling using $p_i = M_i/M$ and $w_{j|i} = c_i$ for all $i = 1, \dots, N$, we have from (3.8) and (3.9) that $\Delta_a = \Delta_b = 0$, that is, there is no difference in the design effects for $\hat{\bar{Y}}$ and \hat{Y} .

The same result as given in example 3.1 can be achieved by $\hat{Y} = M\bar{Y}$. This estimator is the ratio estimator, which can be used if M is known. The case that overall sampling weights are a constant for all the elements (*i.e.*, self-weighting sampling design) is a well known special case. We will come back to this in section 4.

Example 3.2 One-stage simple random cluster sampling or two-stage sample design with srs for both stages. Under these designs, we have $w_{j|i} = c_i$ and $p_i = 1/N$ for all i and j and thus, it follows from (3.8) and (3.9) that $\Delta_b = 0$ and

$$\Delta_a = \begin{cases} 0 & \text{if } M_i = M_0 \text{ for all } i, \\ \bar{m}' \frac{CV_M^2}{CV_y^2} & \text{if } \bar{Y}_i \text{ are all equal,} \\ -\bar{m}' \frac{CV_M^2}{CV_y^2} & \text{if } Y_i \text{ are all equal,} \end{cases} \quad (3.10)$$

where $\bar{m}' = m'/n$, $CV_M^2 = \bar{M}^{-2} \sum_{i=1}^N (M_i - \bar{M})^2/N$ denotes the relative variance of cluster sizes M_i , and $\bar{M} = M/N$ denotes the average size of clusters. The conditions in (3.10) also satisfy the conditions in (3.9) and therefore, (3.10) is a special case of (3.9). Note that the quantity $A_p(y)$ in

expression (3.9) approximately reduces to $\bar{m}' \cdot CV_M^2 / CV_y^2$ when $p_i = 1/N$ for all i .

Example 3.2 shows that when unequal cluster sizes are not reflected in the sampling design, the relative efficiency of \hat{Y} over $\hat{\bar{Y}}$ depends in part on the relative variability of cluster sizes. If the cluster means are all equal, then cluster sampling makes $\hat{\bar{Y}}$ more efficient than \hat{Y} , vice versa if all the cluster totals are equal. On the other hand, if all clusters are equal in size, no difference in the design effects arises by simple random sampling of clusters.

In section 4, we utilize the results derived in this section to discuss other examples used in the sampling literature.

4. EXAMPLES ON THE DESIGN EFFECT IN THE SAMPLING LITERATURE

4.1 Unequal Probability Element Sampling

Consider an unequal probability element sampling design without clustering. The discussion in section 3 applies to this example with $M_i \equiv 1$ for all $i = 1, \dots, N$ and thus, $m = n$. For brevity's sake, we use lower cases y_i to denote the value of the y -variable, and we also assume that N is large so that $N/(N-1) \approx 1$. Due to the absence of the second stage sampling variation, the design effects for \hat{Y} and $\hat{\bar{Y}}$ given in expressions (3.5) and (3.6) reduce to

$$\text{Deft}_p^2(\hat{Y}) \equiv \frac{\sum_{i=1}^N p_i^{-1} (y_i - p_i Y)^2}{\sum_{i=1}^N N (y_i - \bar{Y})^2} \quad (4.1)$$

and

$$\text{Deft}_p^2(\hat{\bar{Y}}) \equiv \frac{\sum_{i=1}^N p_i^{-1} (y_i - \bar{Y})^2}{\sum_{i=1}^N N (y_i - \bar{Y})^2}. \quad (4.2)$$

Further let us consider an example where the survey variable y is not correlated with the selection probability p_i .

Example 4.1 Unequal probability element sampling with no correlation between y_i and p_i . When y_i and p_i are not correlated, we can approximate $\sum_{i=1}^N p_i^{-1} (y_i - \bar{Y})^2$ by $n\bar{W} \sum_{i=1}^N (y_i - \bar{Y})^2$, where $\bar{W} = N^{-1} \sum_{i=1}^N w_i$. Note that $E_p(n^{-1} \sum_{i=1}^N w_i) = N/n$, $E_p(n^{-1} \sum_{i=1}^N w_i^2) = N\bar{W}/n$ and $E_p(n^{-1} \sum_{i=1}^N w_i^2) / E_p(n^{-1} \sum_{i=1}^N w_i) = n\bar{W}/N$. Thus,

$$\begin{aligned} \text{Deft}_p^2(\hat{\bar{Y}}) &\approx n\bar{W}/N \\ &= E_p\left(n^{-1} \sum_{i=1}^N w_i^2\right) / E_p\left(n^{-1} \sum_{i=1}^N w_i\right). \end{aligned} \quad (4.3)$$

It is easy to show that $n\bar{W}/N \geq 1$ using the Cauchy-Schwarz inequality (Apostol 1974, page 14). In addition, routine calculations show from (4.1) and (4.2) that

$$\begin{aligned} & \text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\hat{Y}}) \\ & \equiv \text{CV}_y^{-2} \left\{ \sum_{i=1}^N p_i^{-1} (p_i - \bar{p})^2 - 2Y^{-1} \sum_{i=1}^N p_i (y_i - \bar{Y})(p_i - \bar{p}) \right\} \\ & = \text{CV}_y^{-2} (n\bar{W}/N-1), \end{aligned}$$

where $\bar{p} = N^{-1} \sum_{i=1}^N p_i = 1/N$. The latter expression is obtained from $\sum_{i=1}^N p_i^{-1} (p_i - \bar{p})^2 = n\bar{W}/N-1$ and $\sum_{i=1}^N p_i^{-1} (y_i - \bar{Y})(p_i - \bar{p}) = 0$ because y_i and p_i are uncorrelated. Consequently,

$$\text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\hat{Y}}) \equiv \text{CV}_y^{-2} \left\{ \text{Deft}_p^2(\hat{\hat{Y}}) - 1 \right\}$$

or

$$\text{Deft}_p^2(\hat{Y}) \equiv (1 + \text{CV}_y^{-2}) \text{Deft}_p^2(\hat{\hat{Y}}) - \text{CV}_y^{-2}. \quad (4.4)$$

From (4.4), it is clear that $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\hat{Y}})$ if $\text{Deft}_p^2(\hat{\hat{Y}}) \geq 1$ and the equality holds if $\text{Deft}_p^2(\hat{\hat{Y}}) = 1$ or $\bar{W} = N/n$. Also, $\text{Deft}_p^2(\hat{Y}) < \text{Deft}_p^2(\hat{\hat{Y}})$ if $1/(1 + \text{CV}_y^{-2}) < \text{Deft}_p^2(\hat{\hat{Y}}) < 1$.

Example 4.1 shows that \hat{Y} tends to have a larger design effect than $\hat{\hat{Y}}$ if the correlation between y_i and p_i is weak and $\text{Deft}_p^2(\hat{\hat{Y}}) \geq 1$.

The customary quantification of the effect of unequal weights on the design efficiency shown in (2.2) is due to Kish (1965, 11.7). He considered cases where the unequal weights arise from “haphazard” or “random” sources such as frame problems or non-response adjustments. Assuming that (1) a random sample of size n selected with replacement is divided into G weighting classes such that the same weight w_g is assigned to n_g sampling units within class g and $n = \sum_{g=1}^G n_g$, and that (2) all G weighting class variances are equal to the unit variance of y , i.e., $S_{yg}^2 = S_y^2$ for all $g = 1, \dots, G$, he proposed a quantity given as

$$\text{Deft}_{\text{Kish}}^2(\hat{\hat{Y}}) = n \sum_{g=1}^G n_g w_g^2 / \left(\sum_{g=1}^G n_g w_g \right)^2, \quad (4.5)$$

to measure the increment in the variance of $\hat{\hat{Y}}$ in comparison with the hypothesized variance under srsw of size n . The rationale behind the above derivation is that the loss in precision of $\hat{\hat{Y}}$ due to haphazard unequal weighting can be approximated by the ratio of the variance under disproportionate stratified sampling to that under the proportionate stratified sampling.

In (4.5), letting $n_g = 1$ for all g and thus, $n = G$, Kish (1992) later proposed a well-known approximate formula given as

$$\text{Deft}_{\text{Kish}}^2(\hat{\hat{Y}}) = n \sum_{i=1}^n w_i^2 / \left(\sum_{i=1}^n w_i \right)^2 = 1 + \text{cv}_w^2, \quad (4.6)$$

where $\text{cv}_w^2 = n^{-1} \sum_{i=1}^n (w_i - \bar{w})^2 / \bar{w}^2$ is the sample relative variance and \bar{w} is the sample mean of w_i . Note that (4.6) is a sample approximate of (4.3). For a sampling design

which is inefficient for estimation of Y , the inefficiency diminishes with the ratio estimation. Next, we consider the opposite case where the y -variable is correlated with the selection probability p_i , where the efficiency of \hat{Y} increases.

Example 4.2 Unequal probability element sampling where y_i is correlated with p_i . Suppose that y_i is linearly related with p_i by $y_i = A + Bp_i + e_i$, where A and B are the least-square regression coefficients of the model for the (finite) population and e_i is the corresponding residual. Furthermore, assume that the regression model fits well to the population data and the error variance is roughly homogeneous so that $R_{ew} \equiv 0$ and $R_{e_w}^2 \equiv 0$, where R_{ew} and $R_{e_w}^2$ denote the population correlations of pairs (e_i, w_i) and (e_i^2, w_i) , respectively. For example, $R_{ew} = \sum_{i=1}^N (e_i - \bar{E})(w_i - \bar{W}) / \{ (N-1) S_e S_w \}$, where $\bar{E} = \sum_{i=1}^N e_i / N$; S_e and S_w are the population standard deviations of e_i and w_i , respectively. Then the design effects given by (4.1) and (4.2) reduce to

$$\begin{aligned} \text{Deft}_p^2(\hat{Y}) & \equiv (n\bar{W}/N) (1 - R_{yp}^2) \\ & + (n\bar{W}/N - 1) \left(\frac{R_{yp}}{\text{CV}_p} - \frac{1}{\text{CV}_y} \right)^2 \end{aligned} \quad (4.7)$$

and

$$\begin{aligned} \text{Deft}_p^2(\hat{\hat{Y}}) & \equiv (n\bar{W}/N) (1 - R_{yp}^2) \\ & + (n\bar{W}/N - 1) \left(\frac{R_{yp}}{\text{CV}_p} \right)^2, \end{aligned} \quad (4.8)$$

respectively, where R_{yp} is the population correlation between y_i and p_i and CV_p is the population coefficient of variation of p_i (see Park and Lee (2001) for proof). It follows from (4.7) and (4.8) that $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\hat{Y}})$ if and only if

$$2R_{yp} \leq \text{CV}_p / \text{CV}_y, \quad (4.9)$$

where the equality holds if and only if $2R_{yp} = \text{CV}_p / \text{CV}_y$. Also, the inequality is reversed when the inequality in (4.9) becomes opposite.

The condition (4.9) indicates that \hat{Y} tends to be less efficient in terms of precision than $\hat{\hat{Y}}$ whenever R_{yp} is small. Thus, we see that R_{yp} plays an important role in determining the design efficiency of unequal probability sampling on \hat{Y} and $\hat{\hat{Y}}$ and their relative efficiency.

In an attempt to develop an approximate expression to the design effect when y_i is correlated with p_i , Spencer (2000) proposed a sample approximate formula for \hat{Y} and compared it with Kish's approximate formula (4.6) for the special case of $R_{yp} = 0$. As seen in example 4.2, the two design effects (4.7) and (4.8) are not equal unless $\bar{W} = N/n$ (see Park and Lee (2001) for more discussion

and some numerical examples). In addition, this special case provides the same condition as for example 4.1 and thus, the two approximate design effect formulae (4.7) and (4.8) are equivalent to (4.4) and (4.3), respectively.

4.2 One-Stage Cluster Sampling

Consider a one-stage cluster sampling, where every element in a sampled cluster is included in the sample, *i.e.*, $m_i \equiv M_i$ for all $i \in s_a$. Due to the absence of the second stage sampling variation, the variance of \hat{Y} takes only the first term of expression (3.1) and it can be decomposed as

$$\sum_{i=1}^N w_i (Y_i - p_i Y)^2 = \frac{M(N-1)}{n} S_{yB}^2 + \sum_{i=1}^N w_i Q_i \bar{Y}_i^2, \quad (4.10)$$

where $S_{yB}^2 = (N-1)^{-1} \sum_{i=1}^N M_i (\bar{Y}_i - \bar{Y})^2$ and $Q_i = M_i(M_i - p_i M)$ for $i=1, \dots, N$. Note that $Q_i = 0$ if $p_i = M_i/M$, that is, p_i is proportional to the cluster size M_i . Also, note that S_{yB}^2 is the between-cluster mean square deviation in an analysis of variance. Denoting the within-cluster mean square deviation as $S_{yW}^2 = (M-N)^{-1} \sum_{i=1}^N \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2$, write $S_{yB}^2 = S_y^2 \{1 + \delta(M-N)/(N-1)\}$ with $\delta = 1 - S_{yW}^2/S_y^2$. Since the expected sample size is $m' = n\bar{M}$, the design effect for \hat{Y} can be written from (4.10) as

$$\text{Deft}_p^2(\hat{Y}) = \left(\frac{N-1}{N} \right) \left(1 + \frac{M-N}{N-1} \delta \right) + \frac{n\bar{M}}{CV_y^2} \sum_{i=1}^N \frac{w_i Q_i}{M_i^2} \left(\frac{Y_i}{Y} \right)^2. \quad (4.11)$$

Similarly, the design effect for $\hat{\bar{Y}}$ can be expressed as

$$\text{Deft}_p^2(\hat{\bar{Y}}) \equiv \left(\frac{N-1}{N} \right) \left(1 + \frac{M-N}{N-1} \delta \right) + \frac{n\bar{M}}{CV_y^2} \sum_{i=1}^N \frac{w_i Q_i}{M_i^2} \left(\frac{D_i}{Y} \right)^2. \quad (4.12)$$

We observe that the design effect for $\hat{\bar{Y}}$ differs from that for \hat{Y} in the second term containing $D_i = \sum_{j=1}^{M_i} (y_{ij} - \bar{Y})$ instead of Y_i . In addition, we note that the quantity $\delta = \delta_p(y)$ is the adjusted coefficient of determination (R_{adj}^2) in the regression analysis context. It may be called a homogeneity measure. For more discussion on δ , see Särndal *et al.* (1992, pages 130–131) and Lohr (1999, page 140).

Example 4.3 One-stage simple random sampling of clusters. In this example, if $p_i = 1/N$ for all $i=1, \dots, N$, the two design effects in (4.11) and (4.12) reduce, respectively, to

$$\text{Deft}_p^2(\hat{Y}) = \left(\frac{N-1}{N} \right) \left(1 + \frac{M-N}{N-1} \delta \right) + \frac{1}{N \cdot CV_y^2} \sum_{i=1}^N (M_i - \bar{M}) \left(\frac{M_i}{\bar{M}} \right) \left(\frac{\bar{Y}_i}{\bar{Y}} \right)^2 \quad (4.13)$$

and

$$\text{Deft}_p^2(\hat{\bar{Y}}) \equiv \left(\frac{N-1}{N} \right) \left(1 + \frac{M-N}{N-1} \delta \right) + \frac{1}{N \cdot CV_y^2} \sum_{i=1}^N (M_i - \bar{M}) \left(\frac{M_i}{\bar{M}} \right) \left(\frac{\bar{D}_i}{\bar{Y}} \right)^2, \quad (4.14)$$

where $\bar{M} = M/N$. Since $\text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\bar{Y}}) \propto \sum_{i=1}^N M_i (M_i - \bar{M}) (2\bar{Y}_i - \bar{Y})$, the inequality between design effects for \hat{Y} and $\hat{\bar{Y}}$ depends on the joint distribution of \bar{Y}_i and M_i .

Example 4.4 One-stage simple random sampling of clusters of equal-size. In this case, we have $M_i \equiv M_0$ and $p_i = 1/N$ for all $i=1, \dots, N$ and both design effects in (4.13) and (4.14) can be approximated by the same quantity given as

$$\left(\frac{N-1}{N} \right) \left[1 + \frac{N(M_0-1)}{N-1} \delta \right], \quad (4.15a)$$

since $M_i - \bar{M} = 0$ for all $i=1, \dots, N$.

To introduce the clustering effect on variance estimation, one often uses the simplest form of one-stage simple random cluster sampling as in example 4.4. For example, see Cochran (1977, section 9.4), Lehtonen and Pahkinen (1995, page 91), and Lohr (1999, section 5.2.2). Although these authors adopted a without-replacement sampling scheme, we compare their formulae with our formulae with the with-replacement sampling assumption for the sake of both simplicity and consistency. Furthermore, the comparison is valid because their formulae are defined with the finite population correction incorporated in both numerator and denominator so that its effect is basically cancelled out. Cochran (1977, section 9.4) derived

$$\text{Deft}_p^2(\hat{\bar{Y}}) = \frac{NM_0-1}{M_0(N-1)} [1 + (M_0-1)\rho] \equiv 1 + (M_0-1)\rho, \quad (4.15b)$$

where ρ is called the intracluster correlation coefficient defined by

$$\rho = \frac{2 \sum_{i=1}^N \sum_{j>k=1}^{M_0} (y_{ij} - \bar{Y})(y_{kj} - \bar{Y})}{(M_0-1) \sum_{i=1}^N \sum_{j=1}^{M_0} (y_{ij} - \bar{Y})^2}. \quad (4.15c)$$

Rewriting $\sum_{i=1}^N [\sum_{j=1}^{M_0} (y_{ij} - \bar{Y})]^2 = M_0(N-1)S_{yB}^2$ and $\sum_{i=1}^N \sum_{j=1}^{M_0} (y_{ij} - \bar{Y})^2 = (NM_0 - 1)S_y^2 = (N-1)S_{yB}^2 + N(M_0 - 1)S_{yW}^2$, it is easy to show that

$$\begin{aligned} & 2 \sum_{i=1}^N \sum_{j>k=1}^{M_0} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) \\ &= \sum_{i=1}^N \left[\sum_{j=1}^{M_0} (y_{ij} - \bar{Y}) \right]^2 - \sum_{i=1}^N \sum_{j=1}^{M_0} (y_{ij} - \bar{Y})^2 \\ &= (M_0 - 1) [(NM_0 - 1)S_y^2 - NM_0 S_{yW}^2] \end{aligned}$$

and, thus, from (4.15c), $\rho = 1 - \{NM_0 / (NM_0 - 1)\} (S_{yW}^2 / S_y^2) \cong \delta$ assuming $M_i \equiv M_0$ for all $i = 1, \dots, N$, $NM_0 / (NM_0 - 1) \cong 1$. Therefore, further assuming $(N-1)/N \cong 1$ and $(NM_0 - 1)M_0^{-1}(N-1)^{-1} \cong 1$, both design effect formulae (4.15a) and (4.15b) are approximately equivalent to $1 + (M_0 - 1)\delta$. Other authors arrived at the same approximate formula. This is because δ and ρ essentially measure the same thing, which is the cluster homogeneity. Under this situation, two estimators \hat{Y} and $\hat{\hat{Y}}$ have the same design effect as discussed in example 3.2. Note that this is a simple case of a self-weighting sampling design.

Sämdal *et al.* (1992, section 8.7) compared the design effects for the two estimators under the setting of example 4.3. They also derived a simplified expression $1 + (\bar{M} - 1)\delta$ for (4.13) and (4.14), assuming the covariances of M_i with $M_i \bar{Y}_i^2$ and $M_i \bar{D}_i^2$ are ignorable. Their discussion on the difference between total and mean estimators boils down to Δ_a in example 3.2. They also noted that the design effect can be much more severe for the population total than for the population mean because more is lost through sampling of clusters when the total is estimated than when the mean is estimated.

A common practice to handle unequal cluster sizes is to use a more efficient sampling method that incorporates the size difference such as pps sampling of clusters. Expressions (4.11) and (4.12) can be applied to arbitrary selection probabilities p_i , where p_i are set to be proportional to some size measures $Z_i \geq 0$. The difference between the design effects for \hat{Y} and $\hat{\hat{Y}}$ is explained by Δ_a in (3.9), or alternatively

$$\Delta_a = \frac{m'}{CV_y^2} \sum_{i=1}^N \frac{w_i Q_i}{M^2} \left[\left(\frac{\bar{Y}_i}{\bar{Y}} \right)^2 - \left(\frac{\bar{D}_i}{\bar{Y}} \right)^2 \right]. \quad (4.16)$$

The term Q_i in (4.16) represents the effect of p_i on the variance estimation when size measures other than the actual cluster sizes M_i are used. Thomsen, Tesfu, and Binder (1986) considered the effect of an out-dated size measure among other factors under two-stage sampling with simple random sample of element at the second stage. We will come back to this in section 4.4.

4.3 Self-Weighting Designs

In a self-weighting sample, every sample element has the same weight. This leads to simple forms for both total and mean estimators. They are given by $\hat{Y} = y/f$ and $\hat{\hat{Y}} = y/m$, where $f = m/M$ is the overall sampling fraction and $y = \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij}$ is the sample total. Then just like simple random sampling as shown in (3.4), the two estimators have the same design effect.

A self-weighting sampling design can be implemented in various ways by synchronizing the first stage sampling method with the second stage sampling method (e.g., Kish 1965, section 7.2). For example, if equal probability sampling is used for the first stage sampling, then the second stage should be sampled by an equal probability sampling method with a uniform sampling fraction for all PSUs. As a special case of this, where an srs of PSUs of equal size (i.e., $M_i = M_0$ for all i) is selected, Hansen *et al.* (1953, Vol. II, pages 162–163) showed

$$CV_p^2(\hat{\hat{Y}}) \cong \frac{1}{m} CV_y^2 [1 + \rho(\bar{m} - 1)], \quad (4.17)$$

where $CV_p^2(\hat{\hat{Y}}) = V_p(\hat{\hat{Y}}) / \bar{Y}^2$ is the relative variance of $\hat{\hat{Y}}$ under the sampling design p and ρ is the intraclass correlation coefficient as defined in (4.15c). Since the relative variance of $\hat{\hat{Y}}$ under srswr is $m^{-1} CV_{y\lambda}^2$, the well known approximate design effect formula for $\hat{\hat{Y}}$ under a self-weighting design follows immediately as

$$\text{Def}_p^2(\hat{\hat{Y}}) = 1 + \rho(\bar{m} - 1). \quad (4.18)$$

For one-stage cluster designs, we showed similar forms given in (4.15a) and (4.15b) (see also Yamane 1967, section 8.7). Hansen *et al.* (1953, Vol. II, page 204) further showed $CV_p^2(\hat{Y}) = CV_p^2(\hat{\hat{Y}})$ for a sample design that employs simple random sampling at both stages. This implies that \hat{Y} and $\hat{\hat{Y}}$ have the same design effect.

4.4 Two-Stage Unequal Probability Sampling

Let us first consider the following example.

Example 4.5 A two-stage sampling design where n PSUs are selected with replacement with probability p_i and an equal size simple random sample of $m_0 \geq 2$ elements is selected with replacement from each selected PSU. With routine calculations and simplification, we can show that

$$\text{Def}_p^2(\hat{Y}) \cong 1 + (m_0 - 1)\tau + W_y^*, \quad (4.19)$$

where

$$\tau = \frac{(N-1)S_{yB}^2 + \sum_{i=1}^N (m_0 - 1)^{-1} S_{y_i}^2}{(N-1)S_{yB}^2 + \sum_{i=1}^N (M_i - 1)S_{y_i}^2}, \quad (4.20)$$

$$S_{y_i}^2 = (M_i - 1)^{-1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2, W_y^* = W_y / V_{\text{srswr}}(\hat{Y}_{\text{srs}}) = (m_0 / CV_y^2) \sum_{i=1}^N (Q_i / p_i M^2) (\bar{Y}_i / \bar{Y})^2 (1 + CV_{y_i}^2 / m_0), \quad \text{and}$$

$CV_{yi}^2 = S_{yi}^2 / \bar{Y}_i^2$ denotes the within-cluster relative variance of the y -variable. Similarly,

$$\text{Def}_p^2(\hat{\bar{Y}}) \equiv 1 + (m_0 - 1)\tau + W_d^*, \quad (4.21)$$

where $W_d^* = W_d / V_{\text{srswr}}(\hat{Y}_{\text{srswr}}) = (m_0 / CV_y^2) \sum_{i=1}^N (Q_i / p_i M^2) (\bar{D}_i / \bar{Y})^2 (1 + CV_{di}^2 / m_0)$, and \bar{D}_i and CV_{di}^2 are defined with the transformed variable d ($d_{ij} = y_{ij} - \bar{Y}$) analogously to \bar{Y}_i and CV_{yi}^2 , respectively. (Detailed derivations of expressions (4.19) and (4.21) are available from the authors.) For the case with $m_i = m_0$ for all i , the difference in the design effects given in (4.19) and (4.21) reduces to (3.7) or (4.16). There is no contribution from the second stage sampling to the difference.

Coming back to Thomsen *et al.* (1986) who studied the effect of using an outdated measure of size on the variance, the above discussion on \hat{Y} parallels with their discussion. The only difference is that they assumed a without-replacement sampling scheme at the second stage. Note, however, that the definition of τ in Thomsen *et al.* (1986) is slightly different from (4.20) and from δ in section 4.2. However, there is a close connection between them. To see this, let us write the τ as a function of some quantities b_i 's associated with PSUs as follows:

$$\tau(b_i) = \frac{(N-1)S_{yB}^2 - \sum_{i=1}^N b_i S_{yi}^2}{(N-1)S_{yB}^2 + \sum_{i=1}^N (M_i - 1)S_{yi}^2}.$$

Then the τ in Thomsen *et al.* (1986) is obtained with $b_i = 1$, the τ in example 4.5 with $-1/(m_0 - 1)$, and δ in section 4.2 with $(M_i - 1) / \{\sum_{i=1}^N (M_i - 1) / (N - 1)\}$. Equating Kish's formula (4.18) for $\hat{\bar{Y}}$ to (4.19) for \hat{Y} , they obviously overlooked that the design effects for \hat{Y} and $\hat{\bar{Y}}$ can be very different.

For more general cases, Kish (1987) proposed the following popular formula for $\hat{\bar{Y}}$:

$$\begin{aligned} \text{Def}_{\text{Kish}}^2(\hat{\bar{Y}}) &= \frac{n \sum_{g=1}^G n_g w_g^2}{\left(\sum_{g=1}^G n_g w_g \right)^2} [1 + \rho(\bar{m} - 1)] \\ &= (1 + cv_w^2) [1 + \rho(\bar{m} - 1)]. \end{aligned}$$

This was obtained by applying (4.5) (or (4.6)) and (4.18) recursively to incorporate the effects of both clustering and unequal weights. Gabler, Haeder and Lahiri (1999) justified the above formula for $\hat{\bar{Y}}$ using a superpopulation model defined for the cross-classification of N clusters and G weighting classes. However, the difference between the design effects for $\hat{\bar{Y}}$ and \hat{Y} cannot be exposed by such a model-based approach, since y_k is treated as a random variable while w_k as fixed. Under this approach, $\text{Def}_p^2(\hat{Y})$

differs from $\text{Def}_p^2(\hat{\bar{Y}})$ only by a factor of $(\hat{M} / M)^2$, although the actual difference can be much more pronounced as we have showed in this paper (e.g., expressions (3.7) and (4.23)).

4.5 More General Cases

Weighting survey data involves not only sampling weights but also various weighting adjustments such as post-stratification, raking, and nonresponse compensation. We consider these general cases here.

We can rewrite the first-order Taylor approximation to the weighted mean estimator $\hat{\bar{Y}} = \hat{Y} / \hat{M}$ given in (3.2) as $(\hat{Y} - Y) / Y \approx (\hat{\bar{Y}} - \bar{Y}) / \bar{Y} + (\hat{M} - M) / M$. Taking variance on both sides,

$$\begin{aligned} CV_p^2(\hat{Y}) &\equiv CV_p^2(\hat{\bar{Y}}) + CV_p^2(\hat{M}) \\ &\quad + 2R_p(\hat{\bar{Y}}, \hat{M}) CV_p(\hat{\bar{Y}}) CV_p(\hat{M}), \end{aligned} \quad (4.22)$$

where $CV_p^2(\hat{Y})$, $CV_p^2(\hat{\bar{Y}})$, $CV_p^2(\hat{M})$ are the relative variances of \hat{Y} , $\hat{\bar{Y}}$, and \hat{M} respectively, and $R_p(\hat{\bar{Y}}, \hat{M})$ is the correlation coefficient of $\hat{\bar{Y}}$ and \hat{M} with respect to the complex sampling design p and any weighting adjustments. Since the relative variances of simple sample total and mean \hat{Y}_{srswr} and \bar{y}_{srswr} are $CV_{\text{srswr}}^2(\hat{Y}_{\text{srswr}}) = CV_{\text{srswr}}^2(\bar{y}_{\text{srswr}}) = m^{-1} CV_y^2$ under srswr of size m , it follows from (4.22) that

$$\begin{aligned} \text{Def}_p^2(\hat{Y}) &\equiv \text{Def}_p^2(\hat{\bar{Y}}) \\ &\quad + 2R_p(\hat{\bar{Y}}, \hat{M}) \nabla_p(y) \text{Def}_p(\hat{\bar{Y}}) + \nabla_p^2(y), \end{aligned} \quad (4.23)$$

where $\nabla_p(y) = CV_p(\hat{M}) / CV_{\text{srswr}}(\bar{y}_{\text{srswr}})$ is nonnegative. As an illustration, consider a binary variable y , where $CV_y^2 \equiv (1 - \bar{Y}) / \bar{Y}$ and, thus, $\nabla_p(y)$ can be arbitrarily large as \bar{Y} approaches 1 or small as \bar{Y} approaches zero assuming $CV_p(\hat{M}) \neq 0$. When $\nabla_p(y)$ is near zero, the two design effects are nearly equal. Otherwise, one is larger than the other depending on the values of $\nabla_p(y)$ and $R_p(\hat{\bar{Y}}, \hat{M})$. When the sampling weights are benchmarked to the known population size M , \hat{Y} and $\hat{\bar{Y}}$ have the same design effect since $\hat{M} = M$ and $CV_p(\hat{M}) = 0$. In this case, $\hat{\bar{Y}}$ is not affected by the benchmarking but $\hat{Y} = M \hat{\bar{Y}}$, which is a ratio estimator. Note that poststratification or raking procedures may be used if population size information is available at subpopulation level and we also get equivalent design effects. In general, however, we have $\text{Def}_p^2(\hat{Y}) \geq \text{Def}_p^2(\hat{\bar{Y}})$ if

$$\begin{aligned} R_p(\hat{\bar{Y}}, \hat{M}) &\geq -\frac{1}{2} \frac{\nabla_p(y)}{\text{Def}_p(\hat{\bar{Y}})} \quad \text{or} \\ R_p(\hat{\bar{Y}}, \hat{M}) &\geq -\frac{1}{2} \frac{CV_p(\hat{M})}{CV_p(\hat{\bar{Y}})}, \end{aligned} \quad (4.24)$$

and vice versa.

It is illuminating to look at some specific situations. For example, if $R_p(\hat{Y}, \hat{M}) \geq 0$, then $\text{Def}_p^2(\hat{Y}) > \text{Def}_p^2(\hat{Y})$, however, a negative correlation (i.e., $R_p(\hat{Y}, \hat{M}) < 0$) doesn't necessarily lead to $\text{Def}_p^2(\hat{Y}) \leq \text{Def}_p^2(\hat{Y})$. For a special case of $R_p(\hat{Y}, \hat{M}) = 0$, the difference is given by

$$\text{Def}_p^2(\hat{Y}) - \text{Def}_p^2(\hat{Y}) \equiv \frac{\text{CV}_p^2(\hat{M})}{\text{CV}_{\text{stswr}}^2(\bar{y}_{\text{st}})} \quad (4.25)$$

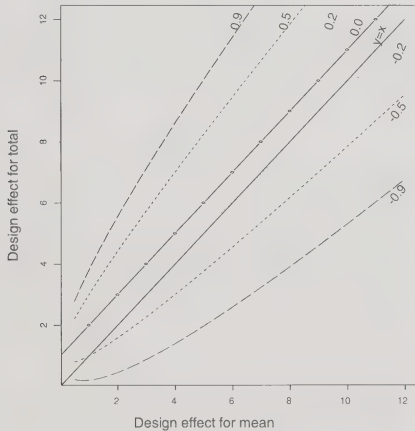
Figure 1 shows graphically the relation between the two design effects. The expression in (4.23) is plotted for some fixed values of $R_p(\hat{Y}, \hat{M})$ and $\nabla_p(y)$. The solid line passing through the origin which represents equal design effects is the reference line. As the graphs show, the comparison is not clear-cut. When $R_p(\hat{Y}, \hat{M}) < 0$, $\text{Def}_p^2(\hat{Y}) \geq \text{Def}_p^2(\hat{Y})$ for small $\text{Def}_p^2(\hat{Y})$ but the relation flips over as $\text{Def}_p^2(\hat{Y})$ grows larger.

Hansen *et al.* (1953, Vol. I, pages 338–339) indicated that $R_p(\hat{Y}, \hat{M})$ would often be close to 0. Under this situation, expression (4.25) is also written as $\text{Def}_p^2(\hat{Y}) \equiv \text{Def}_p^2(\hat{Y}) [1 + \text{CV}_p^2(\hat{M}) / \text{CV}_p^2(\hat{Y})]$, from which we get $\text{Def}_p^2(\hat{Y}) \geq \text{Def}_p^2(\hat{Y})$. This special case was studied by Jang (2001). However, this doesn't seem necessary as can be seen in the following example.

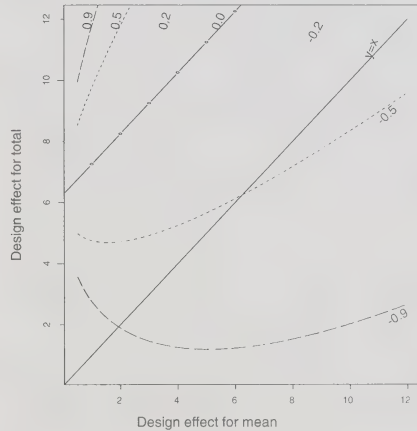
Example 4.6 To illustrate the relationship between the design effects for \hat{Y} and \hat{Y} , we used a data set for the adults collected from the U.S. Third National Health and

Nutrition Examination Survey (NHANES III), which is given as a demo file in WesVar version 4.0. NHANES III is a nationwide large-scale medical examination survey based on a stratified multistage sampling design, for which the Fay's modified balance repeated replication (BRR) method was employed for variance estimation. (See Judkins 1990 for more details on Fay's method.) We used only 19,793 records with complete responses to those characteristics listed in Table 1. Note that the weight in the demo file is different from the NHANES III final weight that was obtained by poststratification. For more detailed information on the demo file, see Westat (2001).

Table 1 presents the design effects for \hat{Y} and \hat{Y} , and component terms of (4.23) for the selected characteristics. Note that $\nabla_p(y)$ monotonically decreases in CV_y given that $m = 19,793$ and $\text{cv}_p(\hat{M}) = 3.2\%$. Although $\nabla_p(y)$ tends to be the determinant factor in the difference of the design effects, $R_p(\hat{Y}, \hat{M})$ can be important when it is negative. For example, for two race/ethnicity characteristics, African American and Hispanic, the negative values, -0.67 and -0.24 of $R_p(\hat{Y}, \hat{M})$ were responsible for $\text{Def}_p^2(\hat{Y}) < \text{Def}_p^2(\hat{Y})$. Some design effects for \hat{Y} are huge. This is not the case with the NHANES III poststratified final weights, with which \hat{Y} and \hat{Y} have the same design effect. This illustrates the importance of benchmarking weight adjustments for total estimates.



(a) $\nabla_p(y) = 1.0$



(b) $\nabla_p(y) = 2.5$

Figure 1. Plots of $\text{Def}_p^2(\hat{Y})$ versus $\text{Def}_p^2(\hat{Y})$ for (a) $\nabla_p(y) = 1.0$ (b) $\nabla_p(y) = 2.5$. The solid line corresponds to $\text{Def}_p^2(\hat{Y}) = \text{Def}_p^2(\hat{Y})$. Other lines correspond to $R_p(\hat{Y}, \hat{M}) = -0.9, -0.5, -0.2, 0, 0.2, 0.5, 0.9$, respectively.

Table 1
Comparison of the design effects for the weighted total and mean using a subset of the adult data file from the U.S. Third National Health and Nutrition Examination Survey (NHANES III)

Characteristic		Mean			Total			cv_y	$r_p(\hat{Y}, \hat{M})$	$\nabla_p(y)$	$-\frac{cv_p(\hat{M})}{2cv_p(\hat{Y})}$
		Estimate	Defl ²	cv	Estimate	Defl ²	cv				
Has smoked 100+ cigarettes in life?	Yes	0.53	4.13	0.014	98,397,795	31.31	0.038	0.944	0.20	4.83	-0.58
Has diabetes?	Yes	0.05	1.75	0.040	9,783,307	1.92	0.042	4.246	-0.34	1.07	-0.31
	No	0.95	1.75	0.002	176,341,218	393.47	0.033	0.236	0.34	19.35	-5.53
Has hypertension/ high blood pressure?	Yes	0.23	3.42	0.024	42,939,866	7.96	0.037	1.826	-0.18	2.50	-0.37
	No	0.77	3.42	0.007	143,184,660	78.44	0.034	0.548	0.18	8.32	-1.22
Race/Ethnicity	African American*	0.12	7.64	0.054	21,567,028	4.21	0.040	2.762	-0.67	1.65	-0.11
	Hispanic*	0.05	6.70	0.079	9,550,326	6.48	0.078	4.300	-0.24	1.06	-0.08
Gender	Male	0.48	1.40	0.009	88,725,967	19.18	0.033	1.048	-0.11	4.35	-1.55
	Female	0.52	1.40	0.008	97,398,559	25.39	0.034	0.954	0.11	4.77	-1.70
Number of cigarettes smoked per day	—	5.25	6.42	0.037	977,225,826	10.51	0.047	2.044	-0.09	2.23	-0.17
Population Size	—	—	—	—	186,124,526	—	0.032	—	—	—	—

Note: * denotes the cases where the design effect for \hat{Y} is smaller than that for \hat{Y} .

5. CONCLUSION

We studied the design effects of the two most widely used estimators for the population mean and total in sample surveys under various with-replacement sampling schemes. We do not think the employment of with-replacement sampling is necessarily a serious limitation because we can see things more clearly without muddling the math with probably unnecessary complications with without-replacement sampling schemes. Furthermore, the effect of the finite population correction is largely canceled out in our formulation of the design effect and so the results are quite comparable with traditional design effects for without-replacement sampling. Therefore, our findings should be useful in practice. We summarize our key findings below.

Kish's well-known approximate formulae for the design effect for (ratio type) weighted mean estimators are not easily generalized in their form and concepts to more general problems, especially weighted total estimators contrary to what many people would perceive. In fact, \hat{Y} and \hat{Y} often have very different design effects unless the sampling design is self-weighting or the sampling weights are benchmarked to the known population size. In addition, the design effect is in general not free from the distribution of the study variable even for the mean estimator, let alone the total estimator. Furthermore, the correlation of the study variable with the weights used in estimation can be an important factor in determining the design effect. Therefore, apart from its original intention, the design effect measures not only the effect of a complex sampling design on a particular statistic but also the effects of the distribution of

the study variable and its relations to the sampling design on the statistic. As complex survey software packages routinely produce the design effect, it seems appropriate to warn the user of the packages of these rather obscure facts about the design effect.

ACKNOWLEDGEMENT

The authors thank Louis Rizza at Westat, an associate editor, and two referees for their helpful comments and suggestions on an earlier version of this paper.

REFERENCES

- APOSTOL, T.M. (1974). *Mathematical Analysis*. 2nd Ed. Reading, MA: Addison-Wesley.
- BARRON, E.W., and FINCH, R.H. (1978). Design Effects in a complex multistage sample: The Survey of Low Income Aged and Disabled (SLIAD), *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 400-405.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd Ed. New York: John Wiley & Sons, Inc.
- CORNFIELD, J. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health*. 41, 654-661.
- GABLER, S., HAEDER, S. and LAHIRI, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*. 25, 105-106.
- HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953). *Sample Survey Methods and Theory*. Vol. I, New York: John Wiley & Sons, Inc.
- HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, Vol. II, New York: John Wiley & Sons, Inc.

- JUDKINS, D.R. (1990). Fay's method for variance estimation, *Journal of Official Statistics*. 6, 223-239.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- KISH, L. (1987). Weighting in Deft². *The Survey Statistician*. June 1987.
- KISH, L. (1992). Weighting for unequal p_i . *Journal of Official Statistics*. 8, 183-200.
- KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*. 11, 55-77.
- JANG, D. (2001). On procedures to summarize variances for survey estimates. *Proceedings of the Survey Research Methods of the American Statistical Association*. In CD-ROM.
- LEHTONEN, R., and PAHKINEN, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.
- LOHR, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks/Cole.
- PARK, I., and LEE, H. (2001). The design effect: do we know all about it? *Proceedings of the Section on Survey Research Methods*, American Statistical Association. In CD-ROM.
- PARK, I., and LEE, H. (2002). A revisit of design effects under unequal probability sampling. *The Survey Statistician*. 46, 23-26.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SPENCER, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology*. 26, 137-138.
- THOMSEN, I., TESFU, D. and BINDER, D.A. (1986). Estimation of Design Effects and Intraclass Correlations When Using Outdated Measures of Size. *International Statistical Review*. 54, 343-349.
- WESTAT (2001). *WesVar 4.0 User's Guide*. Rockville, MD: Westat, Inc.
- YAMANE, T. (1967). *Elementary Sampling Theory*. New Jersey: Prentice-Hall.

Robust Generalized Regression Estimation

JEAN-FRANÇOIS BEAUMONT and ASMA ALAVI¹

ABSTRACT

The Best Linear Unbiased (BLU) estimator (or predictor) of a population total is based on the following two assumptions: i) the estimation model underlying the BLU estimator is correctly specified and ii) the sampling design is ignorable with respect to the estimation model. In this context, an estimator is robust if it stays close to the BLU estimator when both assumptions hold and if it keeps good properties when one or both assumptions are not fully satisfied. Robustness with respect to deviations from assumption (i) is called model robustness while robustness with respect to deviations from assumption (ii) is called design robustness. The Generalized Regression (GREG) estimator is often viewed as being robust since its property of being Asymptotically Design Unbiased (ADU) is not dependent on assumptions (i) and (ii). However, if both assumptions hold, the GREG estimator may be far less efficient than the BLU estimator and, in that sense, it is not robust. The relative inefficiency of the GREG estimator as compared to the BLU estimator is caused by widely dispersed design weights. To obtain a design-robust estimator, we thus propose a compromise between the GREG and the BLU estimators. This compromise also provides some protection against deviations from assumption (i). However, it does not offer any protection against outliers, which can be viewed as a consequence of a model misspecification. To deal with outliers, we use the weighted generalized M -estimation technique to reduce the influence of units with large weighted population residuals. We propose two practical ways of implementing M -estimators for multipurpose surveys; either the weights of influential units are modified and a calibration approach is used to obtain a single set of robust estimation weights or the values of influential units are modified. Some properties of the proposed approach are evaluated in a simulation study using a skewed finite population created from real survey data.

KEY WORDS: Design robustness; Model robustness; M -estimator; Outliers; Shrunk weights; Best linear unbiased predictor.

1. INTRODUCTION

In classical theory, sample data can be viewed as being randomly drawn from an infinite population and assumptions are made about the unknown distribution of the infinite population. In other words, a model is postulated and the interest lies in the estimation of model parameters. In this context, an estimator $\hat{\theta}$ of a model parameter θ is robust if it stays close to the maximum likelihood estimator of θ when the model assumptions hold and if it keeps good properties when the model assumptions are not fully satisfied. The unknown distribution of the infinite population is often assumed to be the normal distribution and, as a result, the maximum likelihood estimator reduces to the usual least-squares estimator.

The presence of outliers in the sample can be viewed as a consequence of a deviation from a model assumption. The majority of the sample could be assumed to come from the selected model but some units, called outliers, could be thought of as coming from a different model. Therefore, the presence of such outliers in the sample may introduce bias and increase the variance of the least-squares estimator of the selected model parameters. Outliers could also be the consequence of a highly skewed distribution. In this case, the least-squares estimator is not biased but may be highly

inefficient due to a deviation from the usual normality assumption. The presence of outliers in the sample could also be the result of measurement errors. However, it is assumed in the rest of this paper that the data have been verified and corrected, if necessary, and that there is no measurement error left in the data. Outlier-robust estimation for infinite populations has been studied extensively (for a review, see Huber 1981; or Hampel, Ronchetti, Rousseeuw and Stahel 1986).

In survey sampling theory, the interest usually lies in the estimation of finite population parameters such as the total, $t_y = \sum_{k \in U} y_k$, of a variable of interest y for a finite population U of size N . Because it is usually not possible to observe the variable y for all population units, the usual practice consists of selecting from the finite population a random sample s of size n according to some probability sampling design $p(s | \mathbf{Z})$. The matrix of design information \mathbf{Z} contains N rows with its k^{th} row equal to \mathbf{z}'_k , and \mathbf{z} is a vector of auxiliary variables available at the design stage. This does not preclude the finite population itself to be assumed to come from a model, as it is explicitly the case when it is chosen to make model-based inferences. Under this type of inference, Royall (1976) derived the Best Linear Unbiased (BLU) estimator (or predictor) \hat{t}_y^B of t_y (see also Valliant, Dorfman and Royall 2000, Chapter 2). It is based

¹ Jean-François Beaumont and Asma Alavi, Household Survey Methods Division, Statistics Canada, 16th floor, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Jean-Francois.Beaumont@statcan.ca and Asma.Alavi@statcan.ca.

on the following two assumptions: i) the estimation model underlying the BLU estimator \hat{t}_y^B is correctly specified and ii) the sampling design is ignorable with respect to the estimation model. In this context, an estimator \hat{t}_y of the finite population total t_y is robust if it stays close to the BLU estimator \hat{t}_y^B when both assumptions hold and if it keeps good properties when one or both assumptions are not fully satisfied. Robustness with respect to deviations from assumption (i) is called model robustness while robustness with respect to deviations from assumption (ii) is called design robustness.

Although we can consider robust estimators that are constructed from a model-based viewpoint, we prefer evaluating their properties as much as possible with respect to the sampling design. This allows us to choose the constants on which robust estimators depend and to evaluate their quality without having to rely on a model and, more specifically, without having to rely on a model for the outliers. This also provides an objective framework for comparing estimators derived under different models. This preference of evaluating properties of model-based estimators with respect to the sampling design is also shared by Little (1983) who notes that design-based asymptotics may be more useful for assessing estimators than model-based asymptotics, particularly when the data set is large.

The Generalized Regression (GREG) estimator of t_y is often viewed as being robust since its property of being Asymptotically Design Unbiased (ADU) is not dependent on assumptions (i) and (ii); that is, the GREG estimator is bias-robust even though its form can be justified by an estimation model. However, if both assumptions hold, the GREG estimator may be far less efficient than the BLU estimator and, in that sense, it is not robust. The relative inefficiency of the GREG estimator as compared to the BLU estimator is caused by widely dispersed design weights. The fact that variable design weights may increase the variance of an estimator is well known (see, for example, Rao 1966; DuMouchel and Duncan 1983; Kish 1992; Pfeffermann 1993; Korn and Graubard 1999, Chapter 4; Elliott and Little 2000; and Kalton and Flores-Cervantes 2003) and is not uncommon in household surveys due to the presence of many weight adjustments before calibration (Kish 1992; and Kalton and Flores-Cervantes 2003). This problem is often treated by truncating the larger design weights (Potter 1988, 1990, 1993; and Stokes 1990).

To obtain a design-robust estimator when the design weights are highly variable, we propose a compromise between the GREG and the BLU estimators based on the weighted Least-Squares (LS) technique. This compromise estimator has a smaller design bias than the BLU estimator when the ignorability assumption is not satisfied and, at the same time, is more efficient than the GREG estimator when

this assumption holds. It also provides some protection against deviations from model assumptions. Balanced sampling (Royall and Herson 1973) and nonparametric calibration (Chambers, Dorfman and Wehrly 1993) are other methods that provide protection against certain types of model misspecifications (see also Valliant, Dorfman and Royall 2000, Chapter 3, 4 and 11). However, none of these methods offer any protection against outliers, which can be viewed as a consequence of a model misspecification. In a model-based framework, the idea underlying the M -estimation technique has been proposed to develop outlier-robust alternatives to the BLU estimator (Chambers 1986; Lee 1991; and Welsh and Ronchetti 1998). In a design-based framework, the M -estimation technique has also been used to develop outlier-robust alternatives to the GREG estimator (Gwet and Rivest 1992; Hülliger 1995 1999; Duchesne 1999; and Zaslavsky, Schenker and Belin 2001). M -estimation is also discussed in the review paper by Lee (1995) and an empirical comparison of several outlier-robust estimators can be found in Gwet and Lee (2000).

Finite population parameters are often very sensitive to the presence of outliers in the population. This is to be contrasted to model (infinite population) parameters, which are usually insensitive to outliers. The problem of outlier robustness is therefore different for finite and infinite populations. As noted in Chambers (1986), it is the sampling error (or the prediction error in a model-based framework) of an estimator which must be insensitive to outliers in finite populations and not necessarily the estimator itself. For instance, when a simple random sampling design is used, the sample median is robust in the classical sense. As a result, its design variance is essentially unaffected by the presence of an outlier in the finite population, no matter how large is that outlier. However, the sampling error and the design bias of the sample median, when used as an estimator of the finite population mean, take an arbitrarily large value when one or more population unit takes an arbitrarily large value. This is explained by the fact that the finite population mean itself takes an arbitrarily large value in such a case. Unlike the sample median, the sample mean is design unbiased but it is not robust in the classical sense. The sampling error and the design variance of the sample mean can thus be very affected by the presence of an outlier in the finite population. This illustrates why outlier-robustness for finite populations is often viewed as a trade-off between bias and variance and why outliers must usually have an influence, at least to some extent, on estimators. The Mean Squared Error (MSE) is therefore a useful criterion for evaluating the quality of outlier-robust estimators of finite population parameters.

The real goal of this paper is to find a robust alternative to the commonly-used GREG estimator of t_y . However, it

is more natural to discuss robustness issues by first introducing the optimal (BLU) estimator. Therefore, the assumptions underlying the BLU estimator are discussed in section 2. We also give additional conditions under which the BLU estimator has a negligible asymptotic design bias. Section 3 deals with design robustness and the weighted LS estimator is introduced. In section 4, model robustness (more specifically, outlier robustness) is discussed and the weighted generalized M -estimation technique is suggested to reduce the influence of units with large weighted population residuals. The proposed estimator is census-consistent in the sense that it is equal to the finite population total t_y when a census is conducted. We propose two practical ways of implementing M -estimators for multipurpose surveys; either the weights of influential units are modified and a calibration approach is used to obtain a single set of robust estimation weights or the values of influential units are modified. Mean Squared Error (MSE) estimation is discussed in section 5. In section 6, some properties of the proposed approach are evaluated in a simulation study using a skewed finite population created from real survey data. Finally, some concluding remarks are made in the last section.

2. THE BEST LINEAR UNBIASED ESTIMATOR

Let us assume that we have a vector of auxiliary variables \mathbf{x} available for all units of the sample s and for which population totals, $\mathbf{t}_x = \sum_{k \in U} \mathbf{x}_k$, are known. Let us also denote by \mathbf{X} , the matrix containing N rows with its k^{th} row equal to \mathbf{x}'_k . The vector \mathbf{x} may or may not contain some variables in the vector \mathbf{z} of design variables. Before discussing robustness, we first describe the two assumptions (see A1 and A2 below) with respect to which robustness is desired. Then, we briefly explain how to validate them.

- A1) The following estimation model m holds: y_k given \mathbf{X} , for $k \in U$, are independently distributed with mean $E_m(y_k | \mathbf{X}) = \mathbf{x}'_k \boldsymbol{\beta}$ and variance $V_m(y_k | \mathbf{X}) = \sigma^2 v_k$, where $\boldsymbol{\beta}$ and σ^2 are unknown model parameters, $v_k = \mathbf{x}'_k \boldsymbol{\lambda}$ and $\boldsymbol{\lambda}$ is a vector of known constants. The subscript " m " indicates that expectations and variances are evaluated with respect to model m .
- A2) The sampling design is independent of \mathbf{y} after conditioning on \mathbf{X} ; that is, $p(s | \mathbf{y}, \mathbf{X}) = p(s | \mathbf{X})$, where \mathbf{y} is a vector containing N elements with its k^{th} element equal to y_k .

Assumption (A1) describes the estimation model m , which specifies the distribution of \mathbf{y} conditional on \mathbf{X} . Standard techniques can be used to validate this model (see, for example, Draper and Smith 1980, Chapter 3). The linearity assumption $E_m(y_k | \mathbf{X}) = \mathbf{x}'_k \boldsymbol{\beta}$ is an important

assumption underlying the estimation model m . There are many ways of assessing the validity of this assumption. A graph of residuals $e_k = y_k - \mathbf{x}'_k \hat{\boldsymbol{\beta}}$ versus $\mathbf{x}'_k \hat{\boldsymbol{\beta}}$, for some m -unbiased estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, is often suggested for this purpose. Any trend in this graph is an indication that the relationship between \mathbf{y} and \mathbf{x} is not linear. To obtain robustness against a deviation from the linearity assumption, a poststratification model can be used when it is possible to partition the population into homogeneous and mutually exclusive groups. An example of the importance of careful modeling in sample surveys can be found in Hedlin, Falvey, Chambers and Kokic (2001).

Assumption (A2) is a sufficient condition for the ignorability (Rubin 1976) of the sampling design with respect to the distribution of \mathbf{y} conditional on \mathbf{X} . In other words, it means that the distribution of \mathbf{y} is independent of s after conditioning on \mathbf{X} . Using assumption (A1), \mathbf{y} can be split into a fixed term $\mathbf{X}\boldsymbol{\beta}$ and a random error term $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Consequently, if the sampling design is independent of $\boldsymbol{\varepsilon}$ after conditioning on \mathbf{X} ; that is, if $p(s | \boldsymbol{\varepsilon}, \mathbf{X}) = p(s | \mathbf{X})$, then assumption (A2) is satisfied and the sampling design is ignorable. Since we only consider sampling designs of the form $p(s | \mathbf{Z})$, an obvious way to make the sampling design ignorable is achieved by including all design variables \mathbf{z} into the estimation model. Examples of such design variables may include the variables used to form the strata, the variable used as a size measure if probability-proportional-to-size sampling is used and so on. The design weights may also provide a useful summary of the design information. Note that it may not be necessary to include all design variables into the estimation model (see Sugden and Smith 1984). Design variables that are independent of \mathbf{y} (or $\boldsymbol{\varepsilon}$) after conditioning on \mathbf{X} should not be included. To assess the validity of assumption (A2), a graph of the residuals, $e_k = y_k - \mathbf{x}'_k \hat{\boldsymbol{\beta}}$, versus design weights w_k (or any design variable) may be useful (see Pfeffermann 1993). Any trend in this graph suggests that the design weights are correlated with the random error $\boldsymbol{\varepsilon}$ and that the sampling design is not ignorable with respect to the estimation model. More formal tests can also be performed to assess the validity of this assumption (see, for example, DuMouchel and Duncan 1983; Graubard and Korn 1993; and, for more references on this topic, Pfeffermann 1993).

Under the estimation model m and the ignorability assumption (A2), it is easy to show that the BLU estimator (Royall 1976) \hat{t}_y^B of t_y takes the simple projection form $\hat{t}_y^B = \mathbf{t}'_x \hat{\mathbf{B}}^B$, where $\hat{\mathbf{B}}^B$ is implicitly defined by the equation

$$\sum_{k \in s} (y_k - \mathbf{x}'_k \hat{\mathbf{B}}^B) \frac{\mathbf{x}_k}{v_k} = \mathbf{0}. \quad (2.1)$$

The BLU estimator can also be written as $\hat{t}_y^B = \sum_{k \in s} w_k^B y_k$, where the BLU estimation weights w_k^B are given by

$$w_k^B = \frac{\mathbf{x}_k'}{v_k} \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{v_k} \right)^{-1} \mathbf{t}_x. \quad (2.2)$$

The model variance $V_m\{(\hat{t}_y^B - t_y) | s, \mathbf{X}\}$ of \hat{t}_y^B is the smallest for every possible sample among all linear m -unbiased estimators of t_y . A direct consequence of this result is that the anticipated variance $E_m\{E_p(\hat{t}_y^B - t_y)^2 | \mathbf{X}\}$ of \hat{t}_y^B is also the smallest among all linear m -unbiased estimators of t_y , where the subscript p indicates that the expectation is evaluated with respect to the sampling design. Under the additional assumption that y_k given \mathbf{X} follows a normal distribution, $\hat{\mathbf{B}}^B$ is also the maximum likelihood estimator of the vector of model parameters $\boldsymbol{\beta}$.

In general, the BLU estimator \hat{t}_y^B is not ADU. However, under the estimation model m , the ignorability assumption (A2) and the following additional assumption (A3), the BLU estimator has the property of being Asymptotically Design Unbiased in Probability (ADUP) in the sense that its relative design bias $E_p(\hat{t}_y^B - t_y)/t_y$ converges in probability to 0 as n and N increase without bound.

A3) $\sum_{k \in U} E_p\{(w_k^B)^2 I_k\} \sigma_k^2 = O(N)$, $\sum_{k \in U} \mathbf{x}_k' \boldsymbol{\beta} = O(N)$ and $\sum_{k \in U} \sigma_k^2 = O(N)$, where $\sigma_k^2 = \sigma^2 v_k$ and I_k is a dummy random variable indicating whether unit k is selected in the sample ($I_k = 1$) or not ($I_k = 0$).

Assumption (A3) describes the asymptotic behaviour of three population quantities. In particular, requiring that $\sum_{k \in U} E_p\{(w_k^B)^2 I_k\} \sigma_k^2 = O(N)$ essentially means that none of the BLU estimation weights becomes too large as the sample size and the population size increase. For instance, if $\mathbf{x}_k = v_k = 1$ and if a sampling design of fixed size n is used, then condition $\sum_{k \in U} E_p\{(w_k^B)^2 I_k\} \sigma_k^2 = O(N)$ is equivalent to assuming that the weights $w_k^B = N/n$ remain bounded as both n and N grow. The proof that \hat{t}_y^B is ADUP is given in the appendix and does not require that $v_k = \mathbf{x}_k' \boldsymbol{\lambda}$. As a result, the BLU estimator is ADUP even when the model variance $V_m(y_k | \mathbf{X})$ is misspecified.

As pointed out above, the BLU estimator is efficient when the estimation model m and the normality assumption hold as well as the ignorability assumption (A2). Under these assumptions and the additional assumption (A3), the BLU estimator is also ADUP. Consequently, a first step towards robustness consists of selecting and validating an estimation model such that these assumptions are satisfied as much as possible. However, they are rarely fully satisfied in practice. For example, one can be reluctant to include all strata identifiers into the estimation model when the number of strata is very large. In such a case, the ignorability assumption might not fully hold. Also, the estimation model, including the normality assumption, may not hold for every variable of interest. Consequently, the non-critical use of the BLU estimator \hat{t}_y^B of t_y is not always appropriate and robust estimators may be needed.

3. DESIGN ROBUSTNESS

Using the fact that $v_k = \mathbf{x}_k' \boldsymbol{\lambda}$, it can be easily shown (see Särndal, Swensson and Wretman, 1992, page 231) that t_y can be expressed as $t_y = \mathbf{t}_x' \mathbf{B}$, where \mathbf{B} is implicitly defined by the equation

$$\sum_{k \in U} (y_k - \mathbf{x}_k' \mathbf{B}) \frac{\mathbf{x}_k}{v_k} = \mathbf{0}. \quad (3.1)$$

The vector \mathbf{B} would be the LS estimator of $\boldsymbol{\beta}$, under the estimation model m , if a census could be conducted. Since \mathbf{t}_x is known, the objective of finding an estimator of the population total t_y is thus equivalent to finding an estimator of \mathbf{B} . In the design-based theory, a natural estimator $\hat{\mathbf{B}}^G$ of \mathbf{B} is implicitly defined by the equation

$$\sum_{k \in s} w_k (y_k - \mathbf{x}_k' \hat{\mathbf{B}}^G) \frac{\mathbf{x}_k}{v_k} = \mathbf{0}, \quad (3.2)$$

where w_k , the design weight of unit k , equals to the inverse of the selection probability π_k . The use of $\hat{\mathbf{B}}^G$ leads to the GREG estimator $\hat{t}_y^G = \mathbf{t}_x' \hat{\mathbf{B}}^G$ of t_y . The GREG estimator \hat{t}_y^G takes a simple projection form because $v_k = \mathbf{x}_k' \boldsymbol{\lambda}$ (see Särndal *et al.* 1992, page 231). It can also be written as $\hat{t}_y^G = \sum_{k \in s} w_k^G y_k$, where the GREG estimation weights w_k^G are given by

$$w_k^G = w_k \frac{\mathbf{x}_k'}{v_k} \left(\sum_{k \in s} w_k \frac{\mathbf{x}_k \mathbf{x}_k'}{v_k} \right)^{-1} \mathbf{t}_x. \quad (3.3)$$

As pointed out in the introduction, the GREG estimator is bias-robust since its property of being ADU is not dependent on the validity of the estimation model m and the ignorability assumption. However, the GREG estimator is not variance-robust since it may be far less efficient than the BLU estimator when both assumptions hold. The inefficiency of the GREG estimator is due to widely dispersed design weights. In household surveys, this situation is not uncommon because of many weight adjustments before calibration. Also, practical considerations for the choice of a sampling design combined with limited information available at the design stage often lead to sampling designs that are approximately ignorable. In household surveys, for instance, geographic information is often the main auxiliary information available to construct the strata. Unless the number of strata is very large, such information is usually weakly correlated with quantitative variables of interest, such as *expenditures* or *income*, and their corresponding population residual variable $E = y - \mathbf{x}' \mathbf{B}$. As a result, the design weight variable w is also weakly correlated with E . This suggests that the ignorability assumption may approximately hold. This also suggests that the design weights act more or less as a random noise when estimating \mathbf{B} using (3.2) and that their influence could be significantly reduced. To obtain a design-robust estimator when the

design weights are highly variable, we thus propose to shrink the design weights towards their mean and to use the LS estimator $\hat{t}_y^{LS} = \mathbf{t}'_x \hat{\mathbf{B}}^{LS}$, where $\hat{\mathbf{B}}^{LS}$ is implicitly defined by

$$\sum_{k \in s} \tilde{w}_k (y_k - \mathbf{x}'_k \hat{\mathbf{B}}^{LS}) \frac{\mathbf{x}_k}{v_k} = \mathbf{0} \quad (3.4)$$

and where \tilde{w}_k is the shrunk weight of unit k given by

$$\tilde{w}_k = \left(\frac{\sum_{k \in s} w_k}{\sum_{k \in s} g(w_k; \alpha)} \right) g(w_k; \alpha). \quad (3.5)$$

The reason for the ratio in the right side of (3.5) is simply to ensure that $\sum_{k \in s} \tilde{w}_k = \sum_{k \in s} w_k$ and the role of the function $g(w_k; \alpha)$ is to obtain shrunk weights \tilde{w}_k that are less variable than the design weights w_k . This function is assumed to be monotone in the constant α , with $1 \leq g(w_k; \alpha) \leq w_k$. The BLU and GREG estimators are therefore extreme special cases of the LS estimator obtained when α is such that $g(w_k; \alpha) = 1$ and $g(w_k; \alpha) = w_k$ respectively. To obtain a simple compromise between these two extreme estimators, we suggest using $g(w_k; \alpha) = w_k^\alpha$, with $0 \leq \alpha \leq 1$. The choice $\alpha = 0$ leads to the BLU estimator while the choice $\alpha = 1$ leads to the GREG estimator. In fact, this suggestion was proposed by Kish (1992, page 198). Other functions $g(w_k; \alpha)$ and other ways of reducing the variability of design weights can be found in the literature (see, for example, Elliott and Little 2000). Truncating large design weights ($g(w_k; \alpha) = \min(w_k, \alpha)$, with $\alpha > 0$) is a common approach that deals with this problem. This approach may be useful when assumptions (A1) and (A2) are not fully satisfied and when there are some abnormally large design weights. A better approach may be to truncate large weighted residuals. The weighted generalized M -estimation technique discussed in the next section can be used for this purpose.

The LS estimator \hat{t}_y^{LS} can also be written as $\hat{t}_y^{LS} = \sum_{k \in s} w_k^{LS} y_k$, where the LS estimation weights w_k^{LS} are given by

$$w_k^{LS} = \tilde{w}_k \frac{\mathbf{x}_k}{v_k} \left(\sum_{k \in s} \tilde{w}_k \frac{\mathbf{x}_k \mathbf{x}_k'}{v_k} \right)^{-1} \mathbf{t}_x. \quad (3.6)$$

Note that the estimation weights w_k^{LS} , including w_k^B and w_k^G as special cases, are calibrated on the known population totals \mathbf{t}_x in the sense that they satisfy the calibration equation $\sum_{k \in s} w_k^{LS} \mathbf{x}_k = \mathbf{t}_x$ (see Deville and Särndal 1992).

4. MODEL (OUTLIER) ROBUSTNESS

As pointed out in the introduction, the LS estimator \hat{t}_y^{LS} provides some protection against deviations from the ignorability assumption and also against deviations from model assumptions. However, it does not offer any

protection against outliers, which can be viewed as a consequence of a model misspecification, including a deviation from the normality assumption. For instance, the GREG estimator is ADU no matter the validity of the estimation model. However, its design variance may be very large in the presence of outliers in the finite population because they may greatly influence its sampling error when they are selected in the sample. This problem may be amplified when the design weights are widely dispersed. For the Horvitz-Thompson estimator, this was well illustrated in the circus example of Basu (1971). Of course, the use of efficient auxiliary variables at the estimation stage can control the impact of outliers on estimates. However, such auxiliary variables are often not available and outlier-robust estimators may provide significant gains over the LS estimator.

Using the Taylor linearization technique (see, for example, Särndal *et al.* 1992, page 235) and given that $t_y = \mathbf{t}'_x \mathbf{B}$, it is well known and easy to show that the sampling error of the GREG estimator can be approximated as follows: $\hat{t}_y^G - t_y \approx \sum_{k \in s} w_k E_k$, where $E_k = y_k - \mathbf{x}'_k \mathbf{B}$ is the population residual for unit k . As a result, a large design weight associated with a large population residual (or outlier) may have a substantial impact on the quality of the GREG estimator. Moreover, it is straightforward to show that the sampling error of the LS estimator can be expressed as $\hat{t}_y^{LS} - t_y = \sum_{k \in s} w_k^{LS} E_k$. Therefore, a large estimation weight associated with a large population residual may greatly influence the sampling error and the quality of the LS estimator. To deal with this problem, we use the Schweppe version (Hampel *et al.* 1986, pages 315 – 316) of the weighted generalized M -estimation technique to reduce the influence of units with large weighted population residuals. This leads to the M -estimator $\hat{\mathbf{B}}^M$ of \mathbf{B} , which is implicitly defined by

$$\sum_{k \in s} \tilde{w}_k \frac{1}{h_k} \psi \left(\frac{h_k \tilde{E}_k(\hat{\mathbf{B}}^M)}{Q} \right) \frac{\mathbf{x}_k}{\sqrt{v_k}} = \mathbf{0}, \quad (4.1)$$

where $\tilde{E}_k(\hat{\mathbf{B}}^M) = (y_k - \mathbf{x}'_k \hat{\mathbf{B}}^M) / \sqrt{v_k}$, Q is a positive population scale parameter and h_k is a weight that may depend not only on \mathbf{x}_k but also on \mathbf{z}_k . The role of the function $\psi(\cdot)$ consists of reducing the influence of units with a large $h_k \tilde{E}_k(\mathbf{B})$. From the above considerations, $h_k = w_k^{LS} \sqrt{v_k}$ or $h_k = \tilde{w}_k \sqrt{v_k}$ is a natural choice. In the former case, the influence of large $w_k^{LS} E_k$ is reduced while, in the latter case, the influence of large $\tilde{w}_k E_k$ is reduced. The choice $h_k = w_k^{LS} \sqrt{v_k}$ may be preferred to $h_k = \tilde{w}_k \sqrt{v_k}$ when there are outliers in the auxiliary variables \mathbf{x} or when α is not close to 1 (assuming $g(w_k; \alpha) = w_k^\alpha$). The main point here is that h_k should depend on survey weights w_k^{LS} or \tilde{w}_k and that both choices suggested above should perform better than simpler choices that do not take into

account the auxiliary variables \mathbf{z} such as $h_k = \sqrt{v_k}$ or $h_k = 1$, which reduce the influence of large unweighted residuals. Also, it should again be noted that the interest is in finding a robust estimator for the vector of population parameters \mathbf{B} and not for the vector of model parameters $\boldsymbol{\beta}$. In fact, \mathbf{B} is itself not robust (in the classical sense) for $\boldsymbol{\beta}$ since it may be highly affected by the presence of outliers in the finite population. As a result, outliers must have a certain influence on $\hat{\mathbf{B}}^M$.

Equation (4.1) can be written in the weighted linear regression form:

$$\sum_{k \in s} \tilde{w}_k^* (\hat{\mathbf{B}}^M, Q) (y_k - \mathbf{x}_k' \hat{\mathbf{B}}^M) \frac{\mathbf{x}_k}{v_k} = 0, \quad (4.2)$$

where

$$\tilde{w}_k^* (\hat{\mathbf{B}}^M, Q) = \tilde{w}_k \frac{\psi(r_k)}{r_k}$$

and

$$r_k = \frac{h_k \tilde{E}_k (\hat{\mathbf{B}}^M)}{Q}.$$

We propose the following modification of the popular function $\psi(\cdot)$ of Huber (1964) that makes the adjusted weights $\tilde{w}_k^* (\hat{\mathbf{B}}^M, Q)$ always greater than or equal to 1: $\psi(r_k) = r_k$, if $|r_k| \leq \varphi$, and $\psi(r_k) = \text{sign}(r_k) \max(|r_k|/\tilde{w}_k, \varphi)$, otherwise, where φ is a positive constant. This leads to adjusted weights

$$\tilde{w}_k^* (\hat{\mathbf{B}}^M, Q) = \begin{cases} \tilde{w}_k, & \text{if } |r_k| \leq \varphi, \\ \max\left(1, \tilde{w}_k \frac{\varphi}{|r_k|}\right), & \text{otherwise.} \end{cases} \quad (4.3)$$

The Iteratively Reweighted Least-Squares (IRLS) algorithm (Beaton and Tukey 1974) is often used to solve (4.2) and (4.3). At a given iteration i , the adjusted weights $\tilde{w}_k^* (\mathbf{B}^{(i-1)}, Q^{(i-1)})$ are first calculated using (4.3) and then $\mathbf{B}^{(i)}$ is obtained by solving (4.2) with $\tilde{w}_k^* (\hat{\mathbf{B}}^M, Q)$ and $\hat{\mathbf{B}}^M$ replaced by $\tilde{w}_k^* (\mathbf{B}^{(i-1)}, Q^{(i-1)})$ and $\mathbf{B}^{(i)}$ respectively. To obtain $\mathbf{B}^{(i)}$, an estimate of Q is usually calculated at each iteration of the IRLS algorithm. In the simulation study of section 6, we have used

$$Q^{(i-1)} = 1.483 \times \text{weighted sample median of } \left(\left| h_k \tilde{E}_k (\mathbf{B}^{(i-1)}) \right| ; k \in s \right), \quad (4.4)$$

where the weighted sample median is calculated using the weights \tilde{w}_k / h_k . Equation (4.4) reduces to the proposal of

Hulliger (1999) when $h_k = 1$ and $g(w_k; \alpha) = w_k$. We suggest using $\mathbf{B}^{(0)} = \hat{\mathbf{B}}^{\text{LS}}$ as the vector of starting values since $\hat{\mathbf{B}}^{\text{LS}}$ is easy to obtain. The iterative procedure is normally repeated until convergence is reached. To reduce computer time, especially if a resampling method is used for MSE estimation, a single iteration of the IRLS algorithm can be performed. In section 6, it is shown empirically that performing a single iteration yields an estimator of the population total that has properties similar to the fully-iterated estimator. This point has also been noted by Lee (1991).

The M -estimator of t_y is given by $\hat{t}_y^M = \mathbf{t}_x' \hat{\mathbf{B}}^M$. With the restriction that $\tilde{w}_k^* (\hat{\mathbf{B}}^M, Q) \geq 1$, where \hat{Q} is an estimator of Q , the estimators $\hat{\mathbf{B}}^M$ and \hat{t}_y^M are census-consistent in the sense that they are exactly equal to \mathbf{B} and t_y respectively, no matter the value of φ and α , when a census is conducted ($\pi_k = 1$, for $k \in U$). This restriction might be useful for controlling the design bias of \hat{t}_y^M when there are shrunk weights \tilde{w}_k close to 1. Note that the estimators $\hat{\mathbf{B}}^M$ and \hat{t}_y^M reduce to $\hat{\mathbf{B}}^{\text{LS}}$ and \hat{t}_y^{LS} respectively when $\varphi = \infty$ ($\psi(r_k) = r_k$). The M -estimator \hat{t}_y^M can also be expressed as $\hat{t}_y^M = \sum_{k \in s} w_k^M y_k$, where the M -estimation weights w_k^M are given by

$$w_k^M = \tilde{w}_k^* (\hat{\mathbf{B}}^M, \hat{Q}) \frac{\mathbf{x}_k}{v_k} \left(\sum_{k \in s} \tilde{w}_k^* (\hat{\mathbf{B}}^M, \hat{Q}) \frac{\mathbf{x}_k \mathbf{x}_k'}{v_k} \right)^{-1} \mathbf{t}_x. \quad (4.5)$$

The estimation weights w_k^M are still calibrated on the known population totals \mathbf{t}_x ($\sum_{k \in s} w_k^M \mathbf{x}_k = \mathbf{t}_x$).

In order to determine appropriate values for α and φ , the MSE of the M -estimator \hat{t}_y^M can be estimated for different choices of α and φ using past or current sample data. Then, the values of α and φ that give the smallest estimated MSE can be chosen. Estimation of MSE is discussed in section 5. As noted in Hulliger (1995), choosing adaptively α and φ by minimizing the estimated MSE with current sample data leads to an estimator \hat{t}_y^M that does not require estimating the scale parameter Q . Also, this procedure controls the magnitude of the design bias of \hat{t}_y^M without requiring the use of additional constants. However, it is likely to provide less efficiency than using the optimal (although unknown) values of α and φ .

In multipurpose surveys, different values of α and φ are likely to be obtained for different variables of interest. If multiple sets of weights are to be avoided, some form of compromise is needed. As a first step towards a compromise, a common value of α , satisfactory for the most important variables of interest, can be determined. Then, we propose two practical ways of implementing the M -estimator \hat{t}_y^M without having to find a compromise value for φ ; either the weights of influential units are modified and a calibration approach is used to obtain a single set of robust estimation weights or the values of influential units

are modified. The former is discussed in section 4.1 while the latter is discussed in section 4.2.

4.1 Modification of the Weights of Influential Units

Let us now assume that it is desired to estimate the population totals of a vector of q variables of interest $\mathbf{y} = (y_1, y_2, \dots, y_q)'$. A vector of q M -estimators $\hat{\mathbf{t}}_y^M = (\hat{t}_{y_1}^M, \hat{t}_{y_2}^M, \dots, \hat{t}_{y_q}^M)'$ of $\mathbf{t}_y = \sum_{k \in U} \mathbf{y}_k$ can be obtained, with potentially different values of φ for different variables. To simplify the notation, we denote the adjusted weights associated with variable y_i by $\tilde{w}_k^*(y_i)$, for $i=1, 2, \dots, q$. Since the adjusted weights $\tilde{w}_k^*(y_i)$ depend on the variable of interest y_i , we obtain q sets of weights, even if a common value of φ is chosen.

Gwet and Rivest (1992), Duchesne (1999) and Hulliger (1999) suggested using the adjusted weights $\tilde{w}_k^*(\mathbf{y}) = \min(\tilde{w}_k^*(y_1), \tilde{w}_k^*(y_2), \dots, \tilde{w}_k^*(y_q))$ to obtain a unique set of weights. Then, estimation weights $w_k^M(\mathbf{y})$ are calculated by replacing $\tilde{w}_k^*(\hat{\mathbf{B}}^M, \hat{Q})$ by $\tilde{w}_k^*(\mathbf{y})$ in (4.5) and \mathbf{t}_y is estimated by $\sum_{k \in s} w_k^M(\mathbf{y}) \mathbf{y}_k$. Although the estimation weights $w_k^M(\mathbf{y})$ are calibrated on the known population totals \mathbf{t}_x , they are not calibrated on the vector of estimates $\hat{\mathbf{t}}_y^M$, which are believed to be our best estimates in the sense of minimizing the estimated MSE. Moreover, the use of $\sum_{k \in s} w_k^M(\mathbf{y}) \mathbf{y}_k$ likely leads to a larger design bias than $\hat{\mathbf{t}}_y^M$ although it controls the design variance. To cope with these issues, we propose computing the estimation weights $w_k^{M,A}(\mathbf{y})$ by replacing $\tilde{w}_k^*(\hat{\mathbf{B}}^M, \hat{Q})$ by the adjusted weights $\tilde{w}_k^*(\mathbf{y})$ in (4.5), and by augmenting the vector of auxiliary variables \mathbf{x} and the known population totals \mathbf{t}_x using \mathbf{y} and $\hat{\mathbf{t}}_y^M$ respectively. As a result, the estimation weights $w_k^{M,A}(\mathbf{y})$ are calibrated on \mathbf{t}_x and $\hat{\mathbf{t}}_y^M$, and \mathbf{t}_y is estimated by $\hat{\mathbf{t}}_y^M = \sum_{k \in s} w_k^{M,A}(\mathbf{y}) \mathbf{y}_k$. Of course, there may be a limit on the number of variables that can be used for calibration purposes. This may somewhat restrict the applicability of this method when q is very large.

4.2 Modification of the Values of Influential Units

Another way of implementing the M -estimator $\hat{\mathbf{t}}_y^M$ in practice consists of modifying the values of the variables of interest \mathbf{y} and using the LS estimation weights w_k^{LS} for all variables. This can be done separately for each variable of interest, so we return to the case of only one variable of interest in this section.

Let us first denote by s_o the random set of all sample units k for which $\tilde{w}_k^*(\hat{\mathbf{B}}^M, \hat{Q}) \neq \tilde{w}_k$. In other words, s_o is the random set of units that have been detected as being influential. Let also $\hat{\mathbf{B}}^{M*}$ be implicitly defined by the equation

$$\sum_{k \in s} \tilde{w}_k(y_{*k} - \mathbf{x}_k' \hat{\mathbf{B}}^{M*}) \frac{\mathbf{x}_k}{v_k} = \mathbf{0}, \quad (4.6)$$

where $y_{*k} = y_k$, if $k \in s - s_o$, and $y_{*k} = y_k^*$, otherwise. The quantity y_k^* is a modified value for the influential unit k that is used to replace y_k . Note that $\hat{\mathbf{B}}^{M*} = \hat{\mathbf{B}}^{LS}$ if $y_{*k} = y_k$, for $k \in s$. The population total t_y can then be estimated by $\hat{t}_y^{M*} = \mathbf{t}_x' \hat{\mathbf{B}}^{M*}$. It is also easy to show that $\hat{t}_y^{M*} = \sum_{k \in s} w_k^{LS} y_{*k}$.

The idea here consists of finding modified values y_k^* , for $k \in s_o$, as close as possible to the original values y_k and that satisfy the constraint $\hat{\mathbf{B}}^{M*} = \hat{\mathbf{B}}^M$. Under this constraint, it is obvious that $\hat{t}_y^{M*} = \hat{t}_y^M$. A possible implementation of this idea is obtained by minimizing the distance function $\sum_{k \in s_o} \tilde{w}_k (y_k - y_k^*)^2 / v_k$ subject to the constraint $\hat{\mathbf{B}}^{M*} = \hat{\mathbf{B}}^M$. This leads to the modified values

$$y_k^* = y_k + \mathbf{x}_k' \left(\sum_{k \in s_o} \frac{\tilde{w}_k}{v_k} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in s} \frac{\tilde{w}_k}{v_k} \mathbf{x}_k \mathbf{x}_k' \right) (\hat{\mathbf{B}}^M - \hat{\mathbf{B}}^{LS}). \quad (4.7)$$

This idea is essentially equivalent to reverse calibration proposed by Ren and Chambers (2002), except that these authors used the constraint $\hat{t}_y^{M*} = \hat{t}_y^M$ instead of $\hat{\mathbf{B}}^{M*} = \hat{\mathbf{B}}^M$. We prefer the latter since it leads to modified values that better preserve the relationships between the variable of interest y and the auxiliary variables \mathbf{x} .

Other ways of determining modified values that satisfy the constraint $\hat{\mathbf{B}}^{M*} = \hat{\mathbf{B}}^M$ can be found. For example, it is straightforward to show that this constraint is satisfied when the following modified values are used:

$$y_k^* = a_k y_k + (1 - a_k) \mathbf{x}_k' \hat{\mathbf{B}}^M, \quad (4.8)$$

where $a_k = \tilde{w}_k^*(\hat{\mathbf{B}}^M, \hat{Q}) / \tilde{w}_k$. The modified values in equation (4.8) have a simple interpretation: they are a weighted average of the robust prediction $\mathbf{x}_k' \hat{\mathbf{B}}^M$ and the observed value y_k . Less weight is given to the observed value y_k when it has a smaller value of a_k and, therefore, when it is highly influential.

5. MEAN SQUARED ERROR ESTIMATION

Estimation of the MSE of \hat{t}_y^M can be used for three different purposes: i) finding appropriate values for α and φ using past or current sample data, ii) evaluating the quality of estimates and iii) making inferences about unknown population quantities. Using the fact that $E_p(\hat{t}_y^G) \approx t_y$, it can be easily shown that the MSE of \hat{t}_y^M can be approximated by

$$\begin{aligned} \text{MSE}_p(\hat{t}_y^M) &\approx V_p(\hat{t}_y^M) \\ &+ E_p(\hat{t}_y^M - \hat{t}_y^G)^2 - V_p(\hat{t}_y^M - \hat{t}_y^G). \end{aligned} \quad (5.1)$$

The last two terms of (5.1) are equal to $[E_p(\hat{\tau}_y^M - \hat{\tau}_y^G)]^2$. They represent the square of the design bias of $\hat{\tau}_y^M$. As suggested in Gwet and Rivest (1992), a potential estimator of $\text{MSE}_p(\hat{\tau}_y^M)$ is given by

$$\text{mse}_p(\hat{\tau}_y^M) = \hat{V}_p(\hat{\tau}_y^M) + \max\left(0, (\hat{\tau}_y^M - \hat{\tau}_y^G)^2 - \hat{V}_p(\hat{\tau}_y^M - \hat{\tau}_y^G)\right), \quad (5.2)$$

where $\hat{V}_p(\hat{\tau}_y^M)$ and $\hat{V}_p(\hat{\tau}_y^M - \hat{\tau}_y^G)$ are estimators of $V_p(\hat{\tau}_y^M)$ and $V_p(\hat{\tau}_y^M - \hat{\tau}_y^G)$ respectively.

Since the estimator $\hat{\tau}_y^M$ has a complex structure, resampling variance estimation methods provide a convenient way of estimating $V_p(\hat{\tau}_y^M)$ and $V_p(\hat{\tau}_y^M - \hat{\tau}_y^G)$. The jackknife, the bootstrap and the balanced repeated replications methods are described and evaluated in Rao, Wu and Yue (1992) for stratified multistage sampling designs, where the primary sampling units are assumed to have been selected with replacement. They have shown in an empirical study that the jackknife variance estimator can have a large bias when estimating the variance of a non-smooth estimator, such as the sample median. Therefore, the jackknife variance estimator might be more biased for estimating the variance of the M -estimator than the balanced repeated replication or the bootstrap method when, at each iteration of the IRLS algorithm, Q is estimated using a non-smooth estimator such as (4.4). Gwet and Lee (2000) studied empirically the performance of the jackknife and the bootstrap methods for some robust estimators. In general, they found encouraging results. It is important to note that the estimator $\hat{\tau}_y^M$ should be recomputed for each resample. This includes repeating the procedure used to estimate α and ϕ if they are estimated using current sample data.

When the goal of MSE estimation is only to find appropriate values for α and ϕ , it may be convenient to consider simplified MSE estimators in order to reduce computer time. We now propose four different ways of simplifying MSE estimation:

- i) Only a single iteration of the IRLS algorithm could be done for each resample even if a fully-iterated M -estimator is used. This might yield reasonable variance estimates since the singly-iterated and fully-iterated M -estimators seem to have similar properties (see section 6.4).
- ii) Some quantities could be assumed fixed (not random) for MSE estimation. This is likely to lead to an underestimation of the MSE but it may be useful if the goal of MSE estimation is only to find appropriate values for α and ϕ . For example, the adjusted weights $\tilde{w}_k^*(\hat{\mathbf{B}}^M, \hat{Q})$ could be assumed fixed. This approximation was in fact suggested in Hulliger (1999). Alternatively, if the M -estimator is implemented using the methodology in section (4.2), the modified values in

(4.7) or (4.8) could be treated as true values for MSE estimation.

- iii) The term $\hat{V}_p(\hat{\tau}_y^M - \hat{\tau}_y^G)$ in (5.2) could be omitted. This would lead to the MSE estimator: $\text{mse}_p(\hat{\tau}_y^M) = \hat{V}_p(\hat{\tau}_y^M) + (\hat{\tau}_y^M - \hat{\tau}_y^G)^2$. Note that this approach leads to an overestimation of the MSE.
- iv) A combination of two of the above three propositions could be considered. For example, the adjusted weights $\tilde{w}_k^*(\hat{\mathbf{B}}^M, \hat{Q})$ could be assumed fixed and the term $\hat{V}_p(\hat{\tau}_y^M - \hat{\tau}_y^G)$ in (5.2) could be omitted. In such a case, an estimator for $V_p(\hat{\tau}_y^M)$ could be obtained by noting that $V_p(\hat{\tau}_y^M) = \mathbf{t}_x' V_p(\hat{\mathbf{B}}^M) \mathbf{t}_x$ and by using the well known Taylor linearization technique of Binder (1983) to estimate $V_p(\hat{\mathbf{B}}^M)$. After some straightforward algebra, we obtain the MSE estimator

$$\begin{aligned} \text{mse}_p(\hat{\tau}_y^M) = & \sum_{k \in s} \sum_{l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} w_k^M (y_k - \mathbf{x}_k' \hat{\mathbf{B}}^M) w_l^M (y_l - \mathbf{x}_l' \hat{\mathbf{B}}^M) \\ & + (\hat{\tau}_y^M - \hat{\tau}_y^G)^2, \end{aligned} \quad (5.3)$$

where π_{kl} is the joint probability of selection of units k and l .

6. SIMULATION STUDY

We performed a simulation study to evaluate some properties of the LS estimator and the M -estimator for a skewed finite population. In particular, we compared a version of the M -estimator that reduces the influence of large weighted population residuals to another one that reduces the influence of large unweighted population residuals. We also compared the performance of the singly- and fully-iterated M -estimators. Section 6.1 describes the population and the sampling design, and sections 6.2 to 6.4 discuss results from the simulation.

6.1 Population and Sampling Design

The data from Statistics Canada's 1998 Survey of Household Spending (SHS) are used to serve as the population. This survey uses a stratified multi-stage design and contains information about 15,457 households on several variables. The variable *Renovation/Repair* is chosen as the variable of interest y . This variable is considered for its greater potential of having very large values. A vector \mathbf{x} of three binary auxiliary variables have been created by dividing the variable *Income* into three categories ($\text{Income} \leq 30,000$, $30,000 < \text{Income} \leq 60,000$ and $\text{Income} > 60,000$) and we have chosen $v_k = 1$, for all $k \in U$. In other words, we have considered a poststratification estimation model, which should give us robustness against deviations from the linearity assumption. The population coefficient of

determination (R^2) for this estimation model is 0.13. This is a typical R^2 in household surveys.

From this population, 5,000 samples of expected sample size 300 have been selected using Poisson sampling. We wanted to give households quite dispersed probabilities of selection resulting in variable design weights. We thus assigned probabilities of selection such that they were proportional to the inverse of the SHS design weights (which include a nonresponse adjustment factor). The selection probabilities are thus given by $\pi_k = (300 / \sum_{k \in U} \pi_k^*) \pi_k^*$, where π_k^* , for $k \in U$, is the reciprocal of the design weight (including a nonresponse adjustment factor) from the SHS data.

Table 6.1 gives some summary statistics for this population. We note that the population residuals are very skewed and that the skewness increases when the residuals are multiplied by the design weights. Figure 6.1 shows a graph of the population residuals versus the design weights. First, we note that there is a clear outlier with a residual greater than 50,000 and with a design weight not close to 1. Fortunately, the most extreme design weights are not associated with large population residuals. Also, although this graph may be misleading because of the huge number of points that are overlapping, there does not seem to be any clear relationship between the population residuals and the design weights. In fact, the coefficient of correlation between the design weights and the population residuals is 0.0049. Such a small coefficient of correlation is not atypical in household surveys, for reasons discussed in section 3, and suggests that the ignorability assumption may hold approximately.

Table 6.1
Summary Statistics about the Population

Variable	Mean	Standard Deviation	Skewness
Renovation/Repair	367	1,124	12.6
Population Residual	0	1,104	12.8
Design Weight	177	170	1.8
Weighted Population Residual	922	295,685	15.0

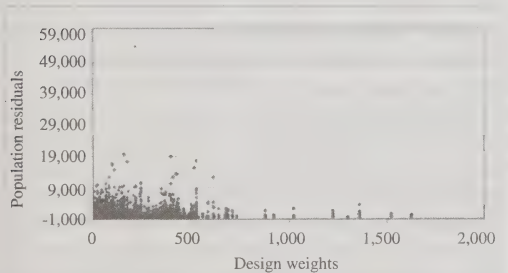


Figure 6.1. Graph of the population residuals versus the design weights

For each of the 5,000 samples, estimates of the population total for the *Renovation/Repair* variable have been calculated for both the LS estimator and two versions of the *M*-estimator; one that reduces the influence of large weighted population residuals ($h_k = \hat{w}_k$) and another one that reduces the influence of large unweighted population residuals ($h_k = 1$). For the i^{th} sample, the relative error in percentage of any estimate \hat{t}_{yi} of t_y is defined as $\Delta_i = 100\% \times (\hat{t}_{yi} - t_y) / t_y$. The Relative Bias (RB) and the Relative Root Mean Squared Error (RRMSE) of any estimator \hat{t}_y , expressed as a percentage of the population total, can thus be estimated by $RB = \sum_{i=1}^{5,000} \Delta_i / 5,000$ and $RRMSE = \sqrt{\sum_{i=1}^{5,000} \Delta_i^2} / 5,000$ respectively. Another measure of interest is the Maximum Absolute Relative Error (MARE) in percentage given by $MARE = \max(|\Delta_i|; i = 1, 2, \dots, 5,000)$. This measure may be useful to assess the sensitivity of an estimator to the presence of influential units in the sample.

6.2 The LS Estimator: Design Robustness

In this section, we evaluate the properties of the LS estimator. Figure 6.2 illustrates the RB, RRMSE and MARE of the LS estimator for 11 values of α ($\alpha = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$) when $g(w_k; \alpha) = w_k^\alpha$. On the one hand, the BLU estimator ($\alpha = 0$) has an RRMSE close to the minimum and the smallest MARE among these 11 values of α but, as expected, leads to the largest RB (in absolute value). Its RB is equal to -13.05% , which is not negligible. Given that a poststratification model is used, this suggests that the ignorability assumption is not fully satisfied even though the correlation between the design weights and the population residuals is small. On the other hand, the GREG estimator ($\alpha = 1$) has a very small RB but has the largest RRMSE and MARE due to the variability of the design weights. When $\alpha = 0.2$, the LS estimator is biased, with an RB of -9.11% , but has a value of MARE relatively close to the smallest value and has the smallest RRMSE (17.94%) among the values of α considered. This is a substantial reduction in comparison with the RRMSE of the GREG estimator (34.77%). In general, values of α between 0.2 and 0.5 provide a reasonable compromise estimator with respect to RB, RRMSE, and MARE. Note that, for larger expected sample sizes, we expect that the minimum MSE be reached for larger values of α because the bias of the LS estimator may dominate its variance.

We have also considered the LS estimator obtained by choosing adaptively, for each selected sample, the value of α that leads to the smallest estimated MSE among the set of 11 values of α considered above. The MSE has been estimated using equation (5.3). The average value of α over the 5,000 selected samples is 0.43. This is slightly larger

than the value of α (0.2) that leads to the smallest MSE (see figure 6.2). This may be due to the simplification made to obtain (5.3), which omits a component of the square design bias when estimating the MSE. Nevertheless, this LS estimator shows a significant improvement over the GREG estimator in terms of RRMSE (26.05%) and MARE (217.99%). This LS estimator shows also a significant improvement over the BLU estimator in terms of RB (−6.24%). Therefore, it seems that choosing adaptively the value of α leads to a useful compromise between the GREG and BLU estimators. However, there is a price to pay in terms of RRMSE by estimating α instead of using the optimal (although unknown) value of α .

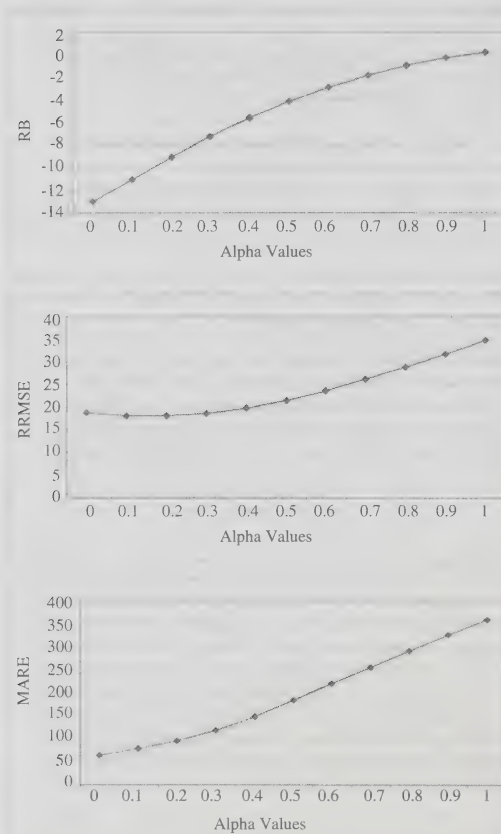


Figure 6.2. RB, RRMSE and MARE of the LS estimator

6.3 The M -estimator: Outlier robustness

We have compared two versions of the M -estimator; one that reduces the influence of large weighted population residuals ($h_k = \tilde{w}_k$) and another one that reduces the

influence of large unweighted population residuals ($h_k = 1$). For the weighted version, we chose 7 values of ϕ ($\phi = 10, 25, 50, 100, 150, 200, \infty$) and for the unweighted version, we chose 9 values of ϕ ($\phi = 2, 5, 8, 11, 14, 17, 20, 30, \infty$). We have only considered the case $\alpha = 1$, as we did not want to confound the effects of changing the constant α with the effect of changing the constant ϕ . Of course, a more efficient estimator could be found by an appropriate choice of both constants. It is to be noted that the results are based on a single iteration of the IRLS algorithm using $\mathbf{B}^{(0)} = \hat{\mathbf{B}}^G$ as the vector of starting values.

It can be seen from figures 6.3 and 6.4 that the weighted version ($h_k = \tilde{w}_k$) has a better potential for reducing the RRMSE and the MARE of M -estimators than the unweighted version ($h_k = 1$). Both graphs of RRMSE present a U -shaped curve. The RRMSE curve for $h_k = \tilde{w}_k$ shows that a value of ϕ between 50 and 150 leads to an RRMSE between 25% and about 27%, while the RRMSE of the GREG estimator (last point on the graphs) is equal to 34.77%. The RRMSE curve for $h_k = 1$ shows that the RRMSE is around 30% for values of ϕ between 8 and 20. In the area where the RRMSE is close to its minimum value, the MARE is smaller when $h_k = \tilde{w}_k$. This suggests that $h_k = \tilde{w}_k$ may control influential units better than $h_k = 1$. As expected, the RB in both figures decreases as ϕ increases.

We have also considered the weighted and unweighted versions of the M -estimator obtained by choosing adaptively, for each selected sample, the value of ϕ that leads to the smallest estimated MSE (using equation 5.3) among the sets of values of ϕ considered above. The average value of ϕ over the selected samples is 72.34 for the weighted version and 10.58 for the unweighted version. Calculation of these averages excludes samples for which $\phi = \infty$ (13 samples for $h_k = \tilde{w}_k$ and 1 sample for $h_k = 1$). Both averages are close to the optimal values of ϕ found in figures 6.3 and 6.4 (100 for $h_k = \tilde{w}_k$, and 11 for $h_k = 1$). The weighted version of the M -estimator has an RB of −10.24%, RRMSE of 28.07% and MARE of 197.86%. The unweighted version of the M -estimator has an RB of −8.26%, RRMSE of 28.18% and MARE of 232.57%. Therefore, both versions of the M -estimator lead to a significant improvement over the GREG estimator in terms of RRMSE and MARE at the expense of an increase in RB (around −10%). The MARE is smaller for the weighted version, which again indicates that it controls influential units better than the unweighted version. However, the difference in the RRMSE between these two estimators is very small. Curiously, it seems that there is no increase in MSE due to estimating ϕ instead of using the optimal value when the unweighted version is used. This observation is somewhat difficult to explain.

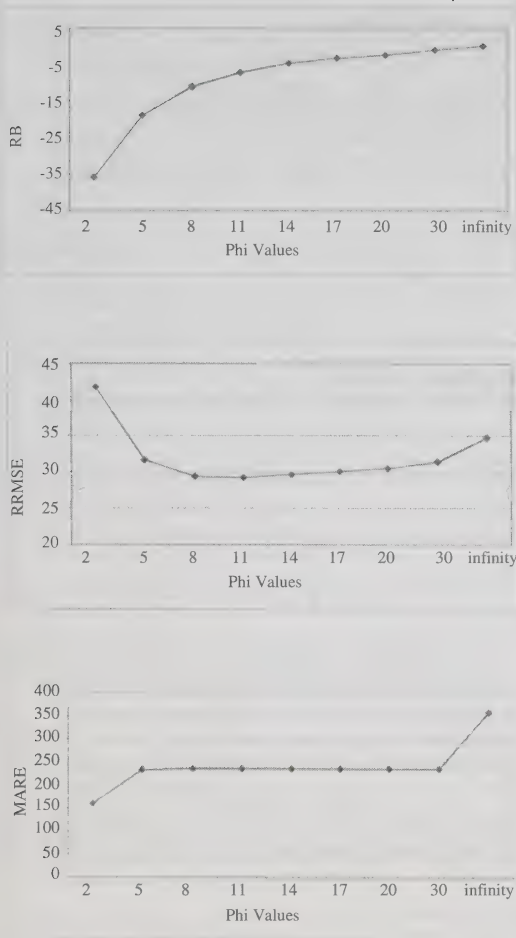


Figure 6.3. RB, RRMSE and MARE of the M -estimator when $h_k = 1$

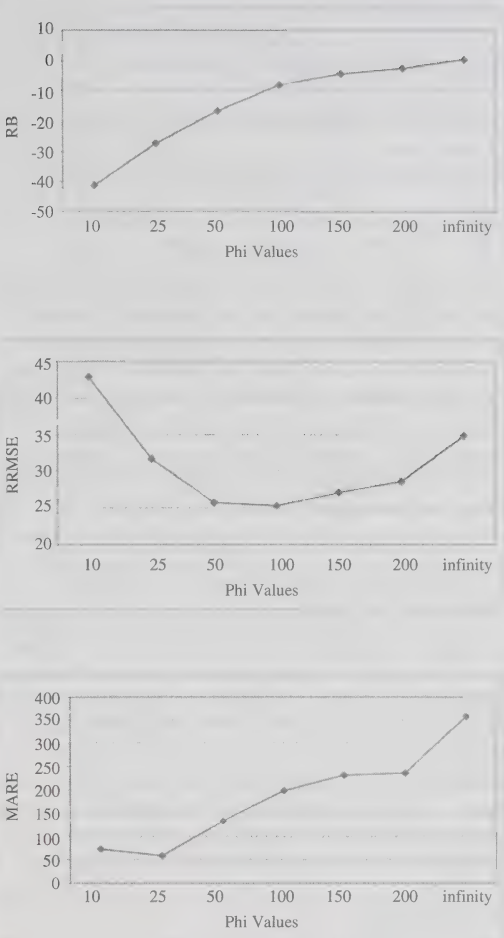


Figure 6.4. RB, RRMSE and MARE of the M -estimator when $h_k = \tilde{w}_k$

6.4 Comparison of the Singly-iterated and Fully-iterated M -estimators

We now compare the singly- and fully-iterated M -estimators when $\alpha = 1$. We only consider the following two cases: i) $h_k = 1$ and $\varphi = 11$; and ii) $h_k = \tilde{w}_k$ and $\varphi = 100$. Most of the time, the IRLS algorithm converged quickly in the fully-iterated case (average number of iterations for convergence is 7.53 for $h_k = 1$, and 7.29 for $h_k = \tilde{w}_k$), but in some of the 5,000 samples (64 for $h_k = 1$, and 75 for $h_k = \tilde{w}_k$) it did not converge. When this situation

occurred, we kept the M -estimate from the last iteration of the IRLS algorithm. From table 6.2, it is evident that the RB, RRMSE and MARE of the singly- and fully-iterated M -estimators are very close to each other. A point worth noting is the slightly smaller RBs for singly-iterated M -estimators. This point has also been observed by Lee (1991) and is likely due to the fact that we used $\mathbf{B}^{(0)} = \hat{\mathbf{B}}^G$ as the vector of starting values for the IRLS algorithm, which is ADU for \mathbf{B} .

Table 6.2
Comparison of Singly- and Fully-iterated M -estimators

Estimator	Singly-iterated			Fully-iterated		
	RB	RRMSE	MARE	RB	RRMSE	MARE
M -estimator ($h_k = 1, \phi = 11$)	-6.94%	29.28%	235.07%	-7.93%	29.27%	235.07%
M -estimator($h_k = \tilde{w}_k, \phi = 100$)	-8.14%	25.36%	197.86%	-8.27%	25.33%	196.73%

7. CONCLUSION

In this paper, we considered robust alternatives to the optimal (BLU) estimator. We first proposed a compromise between the GREG and BLU estimators, the LS estimator, to deal with deviations from the ignorability assumption. The LS estimator is obtained by shrinking the design weights toward their mean. It is expected to be more stable than the GREG estimator when the ignorability assumption holds approximately and less biased than the BLU estimator when this assumption is not fully satisfied. This was confirmed in a simulation study using a population created from real survey data. The LS estimator also offers some protection against deviations from model assumptions.

To deal with outliers, we suggested using the weighted generalized M -estimation technique to reduce the influence of units with large weighted population residuals. We found in a simulation study that significant gains in MSE could be obtained with this method. We also found that an M -estimator obtained using a single iteration of the IRLS algorithm performed similarly to a fully-iterated M -estimator. Finally, we proposed implementing M -estimators for multi-purpose surveys by modifying either the weights of influential units or their values. We believe that both approaches are useful and contribute to bridge a small gap between theory and practice.

ACKNOWLEDGEMENTS

The authors would like to sincerely thank the Associate Editor and three referees for their constructive remarks and suggestions. They would also like to thank Cynthia Bocci and Wesley Yung for their comments, which helped improve the clarity of the paper.

APPENDIX

In this proof, we remove the conditioning on \mathbf{X} when taking expectations and variances with respect to model m in order to simplify the notation. Using Slutsky's theorem, to show that $E_p(\hat{t}_y^B - t_y)/t_y$ converges in probability to 0, as the sample size n and the population size N tend to infinity, under assumptions (A1), (A2) and (A3), it suffices to show that:

$$a) \quad E_p(t_y/\mathbf{t}'_x\boldsymbol{\beta}) = t_y/\mathbf{t}'_x\boldsymbol{\beta} \text{ converges in probability to 1 and}$$

$$b) \quad E_p(\hat{t}_y^B/\mathbf{t}'_x\boldsymbol{\beta}) \text{ converges in probability to 1.}$$

To show (a), note that

$$E_m\left(\frac{t_y}{\mathbf{t}'_x\boldsymbol{\beta}}\right) = 1$$

and

$$V_m\left(\frac{t_y}{\mathbf{t}'_x\boldsymbol{\beta}}\right) = \frac{1}{N} \frac{1}{(\mathbf{t}'_x\boldsymbol{\beta}/N)^2} \sum_{k \in U} \sigma_k^2 / N.$$

By Chebychev's inequality, $t_y/\mathbf{t}'_x\boldsymbol{\beta}$ converges in probability to 1 under model m , as N increases, if $\mathbf{t}'_x\boldsymbol{\beta} = O(N)$ and $\sum_{k \in U} \sigma_k^2 = O(N)$ (assumption A3).

To show (b), we first note that $E_m E_p(\cdot) = E_p E_m(\cdot | s)$ provided that the set of all possible samples does not depend on which population was generated by model m . Consequently, if assumption (A2) holds, it is straightforward to show that $E_m E_p(\hat{t}_y^B/\mathbf{t}'_x\boldsymbol{\beta}) = 1$. Then, we note that

$$V_{mp}\left(\frac{\hat{t}_y^B}{\mathbf{t}'_x\boldsymbol{\beta}}\right) = V_m E_p\left(\frac{\hat{t}_y^B}{\mathbf{t}'_x\boldsymbol{\beta}}\right) + E_m V_p\left(\frac{\hat{t}_y^B}{\mathbf{t}'_x\boldsymbol{\beta}}\right). \quad (\text{A.1})$$

As a result, $V_m E_p(\hat{t}_y^B/\mathbf{t}'_x\boldsymbol{\beta}) \leq V_{mp}(\hat{t}_y^B/\mathbf{t}'_x\boldsymbol{\beta})$ since the two terms on the right side of (A.1) are greater than or equal to 0. By the previous inequality and Chebychev's inequality, $E_p(\hat{t}_y^B/\mathbf{t}'_x\boldsymbol{\beta})$ converges in probability to 1 under model m , as n and N increase, if $\lim_{n, N \rightarrow \infty} V_{mp}(\hat{t}_y^B/\mathbf{t}'_x\boldsymbol{\beta}) = 0$. Using assumption (A2), it is straightforward to show that

$$V_{mp}\left(\frac{\hat{t}_y^B}{\mathbf{t}'_x\boldsymbol{\beta}}\right) = \frac{1}{N} \frac{1}{(\mathbf{t}'_x\boldsymbol{\beta}/N)^2} \sum_{k \in U} E_p\left\{\left(w_k^B\right)^2 I_k\right\} \sigma_k^2 / N.$$

Consequently, $\lim_{n, N \rightarrow \infty} V_{mp}(\hat{t}_y^B/\mathbf{t}'_x\boldsymbol{\beta}) = 0$ if $\mathbf{t}'_x\boldsymbol{\beta} = O(N)$ and $\sum_{k \in U} E_p\{(w_k^B)^2 I_k\} \sigma_k^2 = O(N)$ (assumption A3). This completes the proof.

REFERENCES

- BEATON, A.E., and TUKEY, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147-185.
- BASU, D. (1971). An essay on the logical foundations of survey sampling, part 1. In *Foundations of statistical inference*, (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart, and Winston, 203-233.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- CHAMBERS, R.L., DORFMAN, A.H. and WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DRAPER, N., and SMITH, H. (1980). *Applied regression analysis, second edition*. New-York, John Wiley & Sons, Inc.
- DUCHESNE, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.
- DUMOUCHEL, W.H., and DUNCAN, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- ELLIOTT, M.R., and LITTLE, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.
- GRAUBARD, B.I., and KORN, E.L. (1993). Hypothesis testing with complex survey data: the use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.
- GWET, J.-P., and LEE, H. (2000). An evaluation of outlier-resistant procedures in establishment surveys. In *The Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia, 707-716.
- GWET, J.-P., and RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.
- HAMPEL, F.R., RONCHETTI, E.M., ROUSSEUW, P.J. and STAHEL, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. New-York, John Wiley & Sons, Inc.
- HEDLIN, D., FALVEY, H., CHAMBERS, R. and KOKIC, P. (2001). Does the model matter for GREG estimation? A business survey example. *Journal of Official Statistics*, 17, 527-544.
- HUBER, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.
- HUBER, P.J. (1981). *Robust Statistics*. New-York, John Wiley & Sons, Inc.
- HULLIGER, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.
- HULLIGER, B. (1999). Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 54-63.
- KALTON, G., and FLORES-CERVANTES, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.
- KISH, L. (1992). Weighting for unequal P_i . *Journal of Official Statistics*, 8, 183-200.
- KORN, E.L., and GRAUBARD, B.I. (1999). *Analysis of Health Surveys*. New-York, John Wiley & Sons, Inc.
- LEE, H. (1991). Model-based estimators that are robust to outliers. In *Proceedings of the Annual Research Conference*, Washington, DC, U.S. Bureau of the Census, 178-202.
- LEE, H. (1995). Outliers in business surveys. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott). Chapter 26, New-York, John Wiley & Sons, Inc.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability sampling. *Journal of the American Statistical Association*, 78, 596-604.
- PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- POTTER, F. (1988). Survey of procedures to control extreme sampling weights. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453-458.
- POTTER, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 225-230.
- POTTER, F. (1993). The effect of weight trimming on nonlinear survey estimates. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 758-763.
- RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, Series A, 28, 47-60.
- RAO, J.N.K., WU, C.F.J. and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- REN, R., and CHAMBERS, R.L. (2002). Outlier robust imputation of survey data via reverse calibration. Southampton Statistical Sciences Research Institute Methodology Working Paper M03/19, University of Southampton.
- ROYALL, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- ROYALL, R.M., and HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New-York, Springer-Verlag.
- STOKES, L. (1990). A comparison of truncation and shrinking of sampling weights. In *Proceedings of the 1990 Annual Research Conference*, Washington, DC: Bureau of the Census, 463-471.
- SUGDEN, R.A., and SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

- VALLIANT, R., DORFMAN, A. and ROYALL, R.M. (2000). *Finite population sampling: a prediction approach*. New-York, John Wiley & Sons, Inc.
- WELSH, A.H., and RONCHETTI, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B*, 60, 413-428.
- ZASLAVSKY, A.M., SCHENKER, N. and BELIN, T.R. (2001). Downweighting influential clusters in surveys: application to the 1990 post enumeration survey. *Journal of the American Statistical Association*, 96, 858-869.

Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples

HUI ZHENG and RODERICK J.A. LITTLE¹

ABSTRACT

Samplers often distrust model-based approaches to survey inference because of concerns about misspecification when models are applied to large samples from complex populations. We suggest that the model-based paradigm can work very successfully in survey settings, provided models are chosen that take into account the sample design and avoid strong parametric assumptions. The Horvitz-Thompson (HT) estimator is a simple design-unbiased estimator of the finite population total. From a modeling perspective, the HT estimator performs well when the ratios of the outcome values and the inclusion probabilities are exchangeable. When this assumption is not met, the HT estimator can be very inefficient. In Zheng and Little (2003, 2004) we used penalized splines (*p*-splines) to model smoothly-varying relationships between the outcome and the inclusion probabilities in one-stage probability proportional to size (PPS) samples. We showed that *p*-spline model-based estimators are in general more efficient than the HT estimator, and can provide narrower confidence intervals with close to nominal confidence coverage. In this article, we extend this approach to two-stage sampling designs. We use a *p*-spline based mixed model that fits a nonparametric relationship between the primary sampling unit (PSU) means and a measure of PSU size, and incorporates random effects to model clustering. For variance estimation we consider the empirical Bayes model-based variance, the jackknife and balanced repeated replication (BRR) methods. Simulation studies on simulated data and samples drawn from public use microdata in the 1990 census demonstrate gains for the model-based *p*-spline estimator over the HT estimator and linear model-assisted estimators. Simulations also show the variance estimation methods yield confidence intervals with satisfactory confidence coverage. Interestingly, these gains can be seen for a common equal-probability design, where the first stage selection is PPS and the second stage selection probabilities are proportional to the inverse of the first stage inclusion probabilities, and the HT estimator leads to the unweighted mean. In situations that most favor the HT estimator, the model-based estimators have comparable efficiency.

KEY WORDS: Weighting; REML; Empirical Bayes estimation.

1. INTRODUCTION

In a sample survey, let y_i denote the value of an outcome Y for unit i , and let S denote the set of sampled units. The Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952) $\hat{Y}_{HT} = \sum_{i \in S} y_i / \pi_i$, where π_i is the probability of selection of unit i , is a design-unbiased estimator of the finite population total (and of the mean when divided by the known population count N). It can also be regarded as a model-based projective estimator (Firth and Bennett 1998) for the following linear model relating y_i to π_i :

$$y_i = \beta \pi_i + \pi_i \varepsilon_i,$$

where ε_i is assumed to be i.i.d. normally distributed with mean zero and variance σ^2 .

In Zheng and Little (2003, 2004), we proposed a nonparametric model

$$y_i = f(\pi_i) + \varepsilon_i, \varepsilon_i \sim \text{ind } N(0, \pi_i^{2k} \sigma^2),$$

using penalized splines to model mean of outcome y_i as a smoothly-varying function f of the inclusion probabilities

π_i . We showed in Zheng and Little (2003) that the nonparametric model-based estimators are more efficient than HT for general one-stage probability-proportional-to-size (PPS) samples and not much less efficient than HT when the data are generated using a model that favors HT.

In this article we consider two-stage sampling. In the first stage, a subset of m primary sampling units (PSUs) is drawn from a population with H PSUs with unequal probabilities $\pi_{1,h}$, $h = 1, \dots, H$. Let us number the included PSUs from 1 to m . In the second stage, a simple random sample (srs) of n_h out of N_h secondary sampling units (SSUs) is drawn from the sampled PSU labeled h with probability $\pi_{2,h}$. The overall selection probability for unit i in PSU h is $\pi_h = \pi_{1,h} \pi_{2,h}$, and the HT estimator of the mean of an outcome Y is $\bar{y}_w = \sum_{h=1}^m \sum_{i=1}^{n_h} y_{hi} / (\pi_{1,h} \pi_{2,h}) / N$, where y_{hi} is the value of Y for unit i in PSU h and N is the known total number of units (SSUs) in the whole population. In a commonly adopted design, the first stage selection probability is proportional to an estimate of the PSU size, and the second stage inclusion probabilities are proportional to the inverse of the first stage inclusion probabilities so that the overall inclusion probabilities π_h are equal for all SSUs.

¹ Hui Zheng, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115. E-mail: zheng@hcp.med.harvard.edu; Roderick J.A. Little, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: rlittle@umich.edu.

The inverse probability weighted mean in this case equals the simple sample mean $\bar{y} = \sum_{h=1}^m \sum_{i=1}^{n_h} y_{hi} / \sum_{h=1}^m n_h$.

We assume throughout this article that the selection probabilities $\pi_{1,h}$ are known for all the PSUs $h = 1, \dots, H$. In sections 2 and 3, we assume the PSU counts N_h are also known for all the PSUs in the population. In section 4, we discuss the common situation where N_h is only known for sampled PSUs, but the N_h for nonsampled PSUs can be estimated using a regression model based on auxiliary variables known for all PSUs in the population.

Särndal, Swensson and Wretman (1992) discussed model-assisted alternatives to the HT estimator for two-stage samples with auxiliary information available at the PSU or SSU level. In the first case, let x_h denote a vector of PSU-level auxiliary variables for PSU h . The PSU totals $t_h = \sum_{i=1}^{N_h} y_{hi}$ are assumed to be related to x_h according to a linear model:

$$E(t_h | x_h) = x_h^T \beta, \text{Var}(t_h) = \sigma_h^2, \quad h = 1, \dots, H$$

(Särndal *et al.* 1992). β is estimated by the probability-weighted regression

$$\hat{B} = \left(\sum_{h=1}^m x_h x_h^T / (\sigma_h^2 \pi_{1,h}) \right)^{-1} \sum_{h=1}^m x_h t_h^* / (\sigma_h^2 \pi_{1,h}),$$

where $t_h^* = \sum_{i=1}^{n_h} y_{hi} / \pi_{2,h}$, leading to the projected totals $\hat{t}_h = x_h^T \hat{B}$, $h = 1, \dots, H$. In practice, estimates $\hat{\sigma}_h^2$, either simply assumed (e.g., σ_h proportional to a measure of size of stratum h) or estimated, replace σ_h^2 in the above formula. The generalized regression (GR) estimator of the grand total is

$$\hat{T}_A = \sum_{i=1}^H \hat{t}_i + \sum_{h=1}^m \frac{(t_h^* - \hat{t}_h)}{\pi_{1,h}},$$

and the estimate for the mean is \hat{T}_A / N . The term $\sum_{h=1}^m (t_h^* - \hat{t}_h) / \pi_{1,h}$ is a bias calibration term that makes the estimator design-consistent.

In the second case where auxiliary information is known at the SSU level, let x_{hi} denote the set of auxiliary variables for SSU i in PSU h , $h = 1, \dots, H$; $i = 1, \dots, N_h$. The relationship between the outcome and the auxiliary information is modeled by

$$E(y_{hi} | x_{hi}) = x_{hi}^T \beta, \dots, \text{Var}(y_{hi}) = \sigma_{hi}^2, \quad h = 1, \dots, H, \quad i = 1, \dots, N_h.$$

The probability weighted regression estimate for β is

$$\hat{B} = \left(\sum_{h=1}^m \sum_{i=1}^{n_h} x_{hi} x_{hi}^T / (\sigma_{hi}^2 \pi_{hi}) \right)^{-1} \sum_{h=1}^m \sum_{i=1}^{n_h} x_{hi} y_{hi} / (\sigma_{hi}^2 \pi_{hi}),$$

where π_{hi} is the probability for unit (h, i) to be included in the sample. The GR estimator for the grand total is

$$\hat{T}_B = \sum_{h=1}^H \sum_{i=1}^{N_h} \hat{y}_{hi} + \sum_{h=1}^m \sum_{i=1}^{n_h} \frac{(y_{hi} - \hat{y}_{hi})}{\pi_{hi}},$$

where $\hat{y}_{hi} = x_{hi}^T \hat{B}$. The estimator for the mean is \hat{T}_B / N .

These two methods do not account for the within-PSU correlations of outcome. These correlations can be modeled by treating PSU means as random effects in a hierarchical model. For the case where PSU-level information x_h is available for all PSUs, one such model is:

$$y_{hi} | \mu_h \stackrel{\text{ind}}{\sim} N(\mu_h, \sigma^2) \\ \mu \sim N_H(\varphi, D) \quad (1)$$

where $\mu = (\mu_1, \dots, \mu_H)^T$, $\varphi = (\varphi_1, \dots, \varphi_H)^T$ where μ_h is the mean outcome in PSU h , $\varphi_h = x_h^T \beta$, and D is the covariance matrix of the PSU means. The model-based estimator of \bar{Y} is given by

$$\hat{E}(\bar{Y} | \mathbf{y}, x_h) = \frac{1}{N} \left(\sum_{h=1}^m [n_h \bar{y}_h + (N_h - n_h) \hat{\mu}_h] + \sum_{h=m+1}^H N_h \hat{\mu}_h \right),$$

where $\hat{\mu}_h = \hat{E}(y_{hi} | \mathbf{y}, x_h)$, and \mathbf{y} is the vector of outcomes in the sample.

Different assumptions about φ and D in (1) lead to the following models:

Exchangeable random effects (XRE): (Holt and Smith 1979; Ghosh and Meeden 1986; Little 1991; Lazzaroni and Little 1998): $\varphi_h \equiv \mu_o, h = 1, \dots, H$ and $D = \tau^2 I_H$;

Autoregressive (AR1): (Lazzaroni and Little 1998): $\varphi_h \equiv \mu_o, h = 1, \dots, H$ and $D = r^2 \{\rho^{|h-j|}\}$;

Linear (LIN): (Lazzaroni and Little 1998): $\varphi_h = \alpha + \beta x_h, h = 1, \dots, H$ and $D = \tau^2 I_H$;

Nonparametric: (Elliott and Little 2000): $\varphi_h = f(x_h), h = 1, \dots, H$ and $D = 0$.

The nonparametric models in Elliott and Little (2000) assume nonparametric mean function relating the outcome to the design variables. By assuming $D = 0$, the PSU means are modeled to equal the mean function f instead of varying around it. Nonparametric mixed models relax the assumptions on D (e.g., $D = \tau^2 I_H$) and serve as a natural extension of the Elliott and Little (2000) model and linear mixed models with a parametric mean structure.

It is worth pointing out that some estimators in the above family of models correspond to standard design-based estimators. For example, in an equal-probability design where n_h are approximately constant across PSUs, the unweighted mean corresponds to the special model-based estimator that assumes φ_h is constant.

2. ESTIMATION FOR THE P-SPLINE MIXED MODEL

The linear structure of φ in LIN model is subject to misspecification when the actual mean structure is non-linear. The non-linearity problem can be partially solved by adding polynomial terms (e.g., quadratic or cubic terms) to the fixed effects part in the LIN model. P -spline nonparametric mixed models (Lin and Zhang 1999; Brumback, Ruppert and Wand 1999; Coull, Schwartz and Wand 2001) are even more flexible, since they replace polynomials by smooth nonparametric functions. We propose the following p -spline nonparametric mixed model for inference about the population mean:

P -spline nonparametric mixed model (PMM):

$$\varphi_h = f(x_h), h = 1, \dots, H, D = \tau^2 I_H,$$

where f is a nonparametric degree p spline function:

$$f(x; \beta) = \beta_0 + \sum_{j=1}^p \beta_j x^j + \sum_{l=1}^K \beta_{l+p} (x - \kappa_l)_+^p,$$

where $\kappa_1 < \dots < \kappa_K$ are K fixed knots, $\beta_0, \dots, \beta_{p+K}$ are coefficients to be estimated and $(x)_+^p = x^p \mathbf{1}(x \geq 0)$.

A naive way of estimating $\beta_0, \dots, \beta_{p+K}$ is to treat them as fixed and estimate them together with the variance components σ^2 and τ^2 by fitting a linear mixed model. However this method can yield estimates of f with too much roughness and variability. To avoid overfitting, the roughness of the estimation \hat{f} can be penalized by adding a penalty term to the sum of squared deviations, so that the solution $\hat{\beta}_0, \dots, \hat{\beta}_p$ is minimizes

$$\sum_{h=1}^m (\hat{f}(x_h) - \hat{\mu}_h)^2 + \alpha \sum_{l=1}^K \beta_{l+p}^2.$$

This is achieved in the context of the model by assigning β_0, \dots, β_p flat priors, $(\beta_{p+1}, \dots, \beta_{p+K})$ a normal prior $N_m(0, \sigma_\beta^2)$, and letting $\alpha = \tau^2 / \sigma_\beta^2$. The result is a penalized spline (p -spline) model.

When $p = 1$, \hat{f} is piecewise linear and the coefficients $\beta_0, \dots, \beta_{K+1}$ and σ^2, σ_β^2 and τ^2 are estimated by fitting the linear mixed model:

$$y = X_1 \beta + X_2 u + \varepsilon, \quad (2)$$

where $y = (y_{11}, y_{12}, \dots, y_{mn_m})^T$, $\beta = (\beta_0, \beta_1)^T$, $u = (\beta_2, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$,

$$X_1 = \begin{bmatrix} 1 & x_1 \\ 1 & x_1 \\ \cdot & \cdot \\ \cdot & x_1 \\ \cdot & x_2 \\ \cdot & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_m \end{bmatrix},$$

$$X_2 = \begin{bmatrix} (x_1 - \kappa_1)_+ & \dots & (x_1 - \kappa_K)_+ & 1 & 0 & \dots & 0 \\ \cdot & \dots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \dots & \cdot & \cdot & \cdot & \cdot & \cdot \\ (x_1 - \kappa_1)_+ & \dots & (x_1 - \kappa_K)_+ & 1 & 0 & \dots & \cdot \\ (x_2 - \kappa_1)_+ & \dots & (x_2 - \kappa_K)_+ & 0 & 1 & \dots & \cdot \\ \cdot & \dots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \dots & \cdot & \cdot & \cdot & \cdot & \cdot \\ (x_2 - \kappa_1)_+ & \dots & (x_2 - \kappa_K)_+ & 0 & 1 & \dots & 0 \\ \cdot & \dots & \cdot & 0 & 0 & \dots & 1 \\ \cdot & \dots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \dots & \cdot & \cdot & \cdot & \cdot & \cdot \\ (x_m - \kappa_1)_+ & \dots & (x_m - \kappa_K)_+ & 0 & 0 & \dots & 1 \end{bmatrix},$$

where x_h in X_1 and $(x_h - \kappa_l)_+$ in X_2 are both repeated n_h times. The random terms u and ε are mutually independent with

$$u = (\beta_2, \dots, \beta_{K+1}, u_1, \dots, u_m)^T \sim N_{K+m}(0, G),$$

$$G = \begin{bmatrix} \sigma_\beta^2 I_K & 0 \\ 0 & \tau^2 I_m \end{bmatrix}.$$

Variance components σ^2, σ_β^2 and τ^2 can be estimated by fitting model (2) by restricted maximum likelihood (REML).

The predicted means of PSUs included in the sample are given by: $\hat{\mu} = X_1 \hat{\beta} + X_2 \hat{u}$, where $\hat{\beta} = (X_1^T \hat{V}^{-1} X_1)^{-1} X_1^T \hat{V}^{-1} \bar{y}$, $\hat{u} = \hat{G} X_2^T \hat{V}^{-1} (\bar{y} - X_1 \hat{\beta})$, where $\hat{V} = X_2 \hat{G} X_2^T + \hat{\sigma}^2 \Sigma$, $\Sigma = \text{diag} \{ [1/n_h]_{h=1}^m \}$ and $\bar{y} = (\bar{y}_1, \dots, \bar{y}_m)^T$. The predicted mean for a PSU h that is not selected in the first stage is $\hat{\mu}_h = x_h^T \hat{\beta}^*$, where

$$x_h = [1 \ x_h \ (x_h - \kappa_1)_+ \ \dots \ (x_h - \kappa_K)_+]^T$$

and

$$\hat{\beta}^* = [\hat{\beta}_0 \ \hat{\beta}_1 \ \dots \ \hat{\beta}_{K+1}]^T.$$

Combining the predictions, we obtain the model-based estimator of the population mean

$$\begin{aligned} \hat{E}(\bar{Y} | y, x_h) = \\ \frac{1}{N} \left(\sum_{h=1}^m [n_h \bar{y}_h + (N_h - n_h) \hat{\mu}_h] + \sum_{h=m+1}^H N_h \hat{\mu}_h \right). \end{aligned}$$

3. VARIANCE ESTIMATION METHODS

3.1 Empirical Bayes Model-based Variance

Model (2) can be interpreted as a Bayes model in which the parameters $u = (\beta_2, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$ have multi-variate normal prior $N_{K+m}(0, G)$, and $\beta_0, \beta_1, \sigma^2, \sigma_\beta^2$ and τ^2 all have the flat priors. This leads to the Bayes posterior variance for the vector $(\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$ conditional on σ^2, σ_β^2 and τ^2 as

$$\begin{aligned} \text{Var}((\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T | \sigma^2, \sigma_\beta^2, \tau^2, y) \\ = \sigma^2 (X^T X + \Delta)^{-1} \end{aligned}$$

where $X = [X_1 \ X_2]$ and

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 / \sigma_\beta^2 I_K & 0 \\ 0 & 0 & 0 & \sigma^2 / \tau^2 I_m \end{bmatrix},$$

where I_K and I_m are $(K \times K)$ and $(m \times m)$ identity matrices, respectively.

The empirical Bayes posterior variance for $(\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$ is calculated by replacing σ^2, σ_β^2 and τ^2 with their maximum likelihood (ML) or restricted maximum likelihood (REML) estimates $\hat{\sigma}^2, \hat{\sigma}_\beta^2$ and $\hat{\tau}^2$, respectively. The empirical Bayes method underestimates the true posterior variance, but the underestimation is not severe for the sample sizes encountered in many survey settings. A fully Bayes solution is also possible, but is not covered here.

The predicted population mean is \hat{T}_{pred}/N , where $\hat{T}_{\text{pred}} = T_1 + \hat{T}_2$, $T_1 = \sum_{h=1}^m n_h \bar{y}_h$ is the sample total, and \hat{T}_2 is the estimated total for units not included in the sample,

$$\begin{aligned} \hat{T}_2 = \sum_{h=1}^m (N_h - n_h) \hat{\mu}_h + \sum_{h=m+1}^H N_h \hat{\mu}_h \\ = N_P X_P [\hat{\beta}_0 \ \hat{\beta}_1 \ \dots \ \hat{\beta}_{K+1} \ \hat{\mu}_1 \ \dots \ \hat{\mu}_m]^T, \end{aligned} \quad (3)$$

where

$$N_P = [(N_1 - n_1) \dots (N_m - n_m) N_{m+1} \dots N_H],$$

and

$$X_P = \begin{bmatrix} 1 & x_1 & (x_1 - \kappa_1)_+ & \dots & (x_1 - \kappa_K)_+ & 1 & 0 & \dots & 0 \\ . & . & . & \dots & . & 0 & 1 & 0 & 0 \\ . & . & . & \dots & . & . & . & 0 & . \\ . & . & . & \dots & . & 0 & 0 & 1 & 0 \\ 1 & x_m & (x_m - \kappa_1)_+ & \dots & (x_m - \kappa_K)_+ & 0 & \dots & 0 & 1 \\ 1 & x_{m+1} & (x_{m+1} - \kappa_1)_+ & \dots & (x_{m+1} - \kappa_K)_+ & 0 & \dots & \dots & 0 \\ . & . & . & \dots & . & . & \dots & \dots & . \\ . & . & . & \dots & . & . & \dots & \dots & . \\ . & . & . & \dots & . & . & \dots & \dots & . \\ 1 & x_H & (x_H - \kappa_1)_+ & \dots & (x_H - \kappa_K)_+ & 0 & \dots & \dots & 0 \end{bmatrix}.$$

The empirical Bayes posterior variance for $\hat{\bar{Y}} = \hat{T}_{\text{pred}}/N$ is

$$\begin{aligned} \text{Var}(\hat{\bar{Y}} | \sigma^2, \sigma_\beta^2, \tau^2, X, X_P) = \\ \sigma^2 (N_P X_P (X^T X + \Delta)^{-1} X_P^T N_P^T) / N^2. \end{aligned}$$

3.2 The Jackknife Method

A jackknife variance estimator is developed for the PMM estimator. The jackknife replicates are constructed by dividing the set of PSUs into G equal-sized subgroups and computing the g^{th} pseudo-value as $\hat{\bar{Y}}_g = G\hat{\bar{Y}} - (G-1)\hat{\bar{Y}}_{(g)}$, where $\hat{\bar{Y}}$ is the original PMM estimator and $\hat{\bar{Y}}_{(g)}$ is the same estimator calculated from the reduced sample obtained by excluding the elements from the PSUs in the g^{th} subgroup.

The jackknife variance estimate of $\hat{\bar{Y}}$ is

$$v(\hat{\bar{Y}}) = \frac{1}{G(G-1)} \sum_{g=1}^G (\hat{\bar{Y}}_g - \hat{\bar{Y}})^2,$$

where $\hat{\bar{Y}} = \sum_{g=1}^G \hat{\bar{Y}}_g / G$. In order to balance the distribution of the selection probabilities across the subgroups, sampled units are stratified into n/G strata each of size G with similar first stage inclusion probabilities, and the G subgroups are constructed by randomly selecting one element from each stratum. To save computation, estimates $\hat{\sigma}^2, \hat{\sigma}_\beta^2$ and $\hat{\tau}^2$ are not recomputed for each replicate. That is, we compute pseudo-values of $(\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$ based on the variance components estimated from the whole sample.

Miller (1974) and Shao and Wu (1987, 1989) proved asymptotic properties of the jackknife estimator and jackknife variance estimation in the case of multiple linear

regression. Zheng and Little (2004) provided a theoretical justification for the jackknife method for the p -spline model-based estimator in the case of one-stage designs. Numerical simulations in section 4 suggest the above described jackknife method also works well for the two-stage design. Improved performance might be achieved using the weighted jackknife proposed by Hinkley (1977).

3.3 The Balanced Repeated Replication Method

The BRR method can be applied in stratified designs with two units sampled in each stratum. For designs with one PSU per stratum, strata are often collapsed (Kaltan 1977) for BRR variance estimation. In our application we assume the PSUs are sampled systematically from a randomly ordered list. This can be viewed approximately as a stratified design with n strata each consisting of PSUs with cumulative measures of approximate size $\sum_{i=1}^H z_i / n$, where z_i are the measures of size for the PSUs. One PSU is sampled from each of the n strata. Assuming n is even, the design can be approximated by a stratified design with $n/2$ strata with measures of size $2\sum_{i=1}^N z_i / n$, and two units sampled per stratum. Balanced repeated half samples are constructed by selecting one PSU from each stratum, with the selection scheme based on Hadamard matrices (Plackett and Burman 1946). Let \hat{Y}_b be the p -spline estimator computed from the b^{th} half sample, using the same knots as used in the computation using the full sample – the number and placement of knots needs to allow the spline model to be fitted on each half-sample. The BRR estimator is given by $v_{\text{BRR}}(\bar{Y}) = 1/B \sum_{b=1}^B (\bar{Y}_b - \bar{Y})^2$. This estimate of the variance is subject to some bias, because it treats the design as if it was stratified with two PSUs per stratum.

4. WHEN SOME PSU COUNTS ARE NOT KNOWN

In sections 2 and 3 we assumed that the PSU counts N_h are known for sampled and non-sampled PSUs. In this section we discuss the situation where N_h is only known exactly for the sampled PSUs (labeled 1 through m). We also assume that values $M_h, h = 1, \dots, H$ of an auxiliary variable predictive of N_h are known for the whole population. For example, the M_h may be PSU counts estimated from outside sources such as a census. We conduct a regression of N_h on M_h using the sampled PSUs and replace the counts N_h in (3) for nonsampled PSUs with predictions $\hat{N}_h, h = m+1, \dots, H$ from this regression. The resulting estimate of the total is

$$\tilde{T} = T_1 + \sum_{h=1}^m (N_h - n_h) \hat{\mu}_h + \sum_{h=m+1}^H \hat{N}_h \hat{\mu}_h.$$

The variance estimate of \tilde{T} needs to incorporate the additional variability in \hat{N}_h . In particular, a model-based variance for \tilde{T} is

$$\begin{aligned} \text{Var}(\tilde{T} \mid \pi_h, M_h) &= \text{Var}(E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)) \\ &\quad + E(\text{Var}(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)), \end{aligned}$$

where

$$E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h) = \sum_{h=1}^m (N_h - n_h) \mu_h + \sum_{h=m+1}^H \hat{N}_h \mu_h$$

and

$$\text{Var}(\tilde{T} \mid \hat{N}_h, \pi_h, M_h) \approx \sigma^2 (\tilde{N}_p X_p (X^T X + \Delta)^{-1} X_p^T \tilde{N}_p^T),$$

$\tilde{N}_p = [(N_1 - n_1) \dots (N_m - n_m) \hat{N}_{m+1} \dots \hat{N}_H]$, and X , X_p and Δ are defined as in (3).

If the models for μ_h and N_h are both correctly specified, the above variance can be estimated according to the corresponding models.

5. SIMULATIONS

5.1 Simulation Design

Two simulations are conducted to compare the inverse probability weighting method, the model-assisted method (Särndal *et al.* 1992) and the PMM method in the case of two-stage samples.

In our first simulation, artificial populations are generated with different mean functions $f(\pi_{1,h})$ of the first stage inclusion probabilities. Four different mean functions are simulated: 1) NULL, a constant function; 2) LINDOWN, a linearly decreasing function; 3) EXP, an exponentially increasing function; and 4) SINE, a sine function.

Two combinations of values for variance components are simulated: 1) $\sigma = 0.1$ and $\tau = 0.2$; 2) $\sigma = 0.2$ and $\tau = 0.1$. Only normal errors around the mean functions are simulated while both normal and lognormal within-PSU errors are simulated.

The population consists of 500 PSUs, and in the first stage 48 PSUs are sampled systematically with probability proportional to size (PPS) from a randomly-ordered list. The PSU sizes are uniformly distributed with values ranging from 4 to about 400. The SSU count in each PSU is generated from a distribution with mean equal to 1.05 times the measure of size and log-normal errors with standard deviation 30.

Two types of second-stage sampling plans are studied: 1) within-PSU simple random sampling (srs) with inclusion probabilities proportional to the inverse of the first stage inclusion probabilities, resulting in an equal inclusion probability for all SSUs; 2) within-PSU simple random sampling with the same sampling rate across sampled PSUs, so that the resulting inclusion probabilities for the SSUs in PSU h are proportional to $\pi_{1,h}$.

For each sample drawn under both sampling plans, the following methods are applied:

- A. The HT estimator.
- B. The model-assisted estimation method. We use a linear model regressing the outcome y_{hi} on the first stage inclusion probabilities, which are treated as element-level information. The GR estimator is computed by the formula given in section 1.
- C. The PMM method, with the first-stage inclusion probabilities $\pi_{1,h}$ as the covariate. We use 20 equal percentiles of $\pi_{1,h}$ of the sampled PSUs as the knots for p -spline regression.
- D. The PMM method with the PSU means μ_h estimated the same way as in C, but using estimated PSU counts from a simple linear regression of N_h on the measures of size, which are proportional to $\pi_{1,h}$. This part of the simulation is conducted to study the method described in section 4.

Estimates of \bar{Y} from methods A-D are calculated for each of the 500 samples drawn repeatedly from the artificial populations (each artificial population is generated only once). For the PMM estimator, we compute the empirical Bayes, the jackknife ($K=8$) and BRR variance estimators for each repeated sample. The mean estimate for the

variance of PMM and the coverage rate of the corresponding 95% confidence interval are used to judge the quality of inference. For method D, we study the empirical bias of the model-based variance estimator described in section 4, together with coverage rates of associated confidence intervals.

In the second simulation study, we draw samples of household income data from the 5% public use microdata sample (PUMS) for the State of Michigan in the 1990 US Census, which we treat as a finite population. This simulation is more realistic than the previous simulation in that the outcome values are drawn from a real rather than simulated distribution. The PSUs we simulate are based on the natural geographical clusters called "Public Use Microdata Areas" (PUMAs), which are typically counties and places. There are 67 PUMAs in the Michigan 5% PUMS, with counts of families ranging from around 1,300 to over 10,000. We increase the number of available PSUs by dividing each PUMA into 5, resulting in 335 PSUs. The PSU count ranges from 134 to 3,058. Figure 1 gives the scatter plot of one sample of the average household income versus sampled PSU sizes together with the regression curve $\hat{f}(x)$.

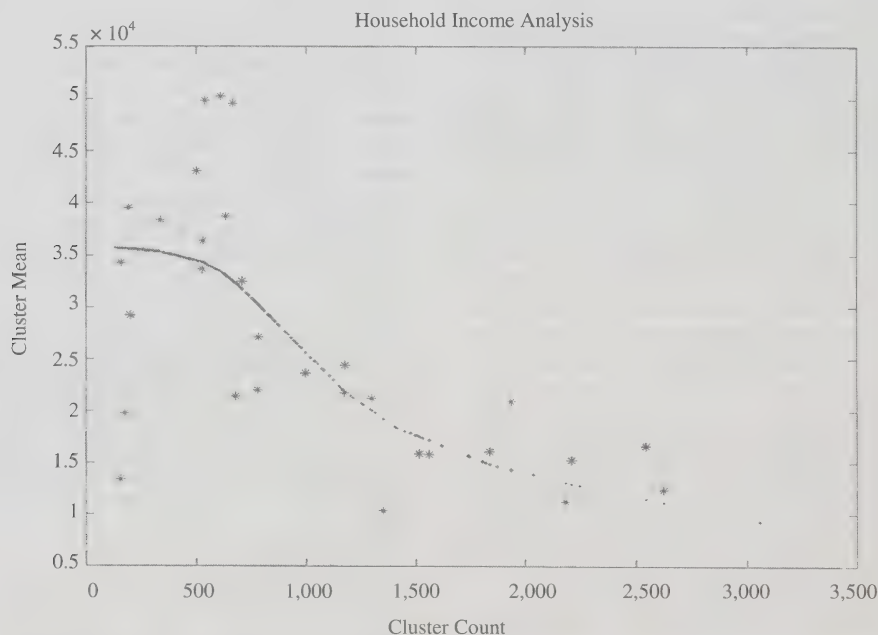


Figure 1. P -spline Regression Curve (dotted line) and the Average Household Income (stars) in Sampled PSUs

Five hundred two-stage samples are drawn, each consisting of 30 PSUs and 20 SSUs (families) from each selected PSU. The first stage sampling is systematic PPS where the measures of size are equal to the PSU counts. The second stage sample is simple random sampling with inclusion probabilities proportional to the inverse of the first stage inclusion probabilities. In the estimation of the mean, we use the true PSU counts as variable x_h , with values proportional to the first-stage inclusion probabilities. We apply the p -spline nonparametric mixed model formulated in (2). We use 10 equally spaced sample percentiles of the PSU counts as the knots in the p -spline.

5.2 Results

Table 1 gives the empirical bias and root mean squared error (RMSE) from four estimation methods of the finite population mean applied to equal probability sample from populations generated with both normal and log-normal within-PSU errors and two (σ, τ) combinations. The empirical bias and RMSE are estimated by the mean bias and squared error from the 500 repeated samples.

Table 1 suggests the PMM based methods give estimators with small biases. In the case of equal probability sampling, the PMM estimator is roughly as efficient as HT estimator when the mean function f is constant. In the more general cases such as NULL and LINDOWN, where f is linear but not constant, the linear model-assisted and PMM method are comparable and both are more efficient than the HT estimator in terms of root mean squared error. For populations EXP and SINE, whose mean functions are

not linear, the PMM method is superior to both the HT and the linear model-assisted estimators. The improvement of efficiency requires the knowledge of complete design information including probabilities $\pi_{1,h}$ and PSU counts N_h for the whole population. When using estimated PSU counts \hat{N}_h in the place of N_h , the resulting estimator is less efficient than in the case with known N_h , but the PMM estimator can still outperform the HT when the mean function is non-constant. Comparisons on populations with normal or log-normal within-PSU errors result in similar findings.

Similar gains for the PMM method are seen in Table 2, for the case of unequal probability sampling. This suggests that the key to improved efficiency is the better prediction given by the nonparametric models. Tables 1 and 2 both suggest that the p -spline model-based estimators have very small empirical design-biases. We believe this is because the flexible mean functions yield good predictions of the PSU means.

Table 3 compares point estimation and coverage of 95% confidence intervals from three variance estimation methods for PMM: the empirical Bayes model-based method, the Jackknife method and the BRR method. The empirical Bayes method is generally satisfactory but tends to underestimate the true variance of PMM estimator, resulting in under-coverage in some cases. The jackknife and the BRR methods tend to yield more robust estimates for the variance. In general, PMM yields estimates with improved efficiency over the traditional HT and linear model-assisted estimators and satisfactory design-based inferences.

Table 1
Empirical Biases and RMSE of PMM, HT, GR and PMM with Estimated N_h for Samples Under Equal Probability Designs

		PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Normal	($\times 10^{-3}$)								
	NULL	1.1	29.7	0.8	30.0	0.8	29.9	1.3	30.1
	Errors								
	LINDOWN	3.5	30.7	3.6	36.4	3.7	30.7	2.3	30.4
$\tau = 0.2$	EXP	-4.4	29.1	-9.4	53.0	-9.5	36.7	-4.3	29.1
	$\sigma = 0.1$								
	SINE	4.8	32.5	2.1	42.0	-0.3	35.9	5.2	34.3
Normal	NULL	5.7	22.0	6.6	22.5	6.6	22.1	5.5	22.3
	Errors								
	LINDOWN	0.5	20.4	-0.6	27.1	-0.3	20.5	1.6	20.6
	$\tau = 0.1$								
	EXP	0.9	23.1	1.9	50.3	-4.2	31.7	0.4	23.4
$\sigma = 0.2$									
	SINE	7.0	22.3	6.5	34.9	3.8	26.4	8.0	26.4
Log-normal	NULL	1.7	32.3	0.9	32.3	0.7	32.3	1.5	32.5
	Errors								
	LINDOWN	2.9	31.9	3.8	39.4	2.7	32.1	3.2	32.0
	$\tau = 0.2$								
	EXP	-0.6	28.4	-5.9	51.5	-6.9	36.4	-0.3	28.5
$\sigma = 0.1$									
	SINE	6.9	33.8	1.5	43.7	-1.9	39.0	-3.1	35.0
Log-normal	NULL	8.5	30.5	9.6	31.3	9.2	31.0	9.1	30.8
	Errors								
	LINDOWN	3.6	32.3	1.9	37.5	3.6	32.1	6.4	33.1
	$\tau = 0.1$								
	EXP	3.9	29.0	6.8	53.8	1.0	34.4	3.7	29.4
$\sigma = 0.2$									
	SINE	-2.9	30.1	-8.9	44.7	-12.0	38.4	-3.8	35.9

Table 2

Empirical Biases and RMSE of PMM, HT, GR and PMM with Estimated N_h for Samples Under Unequal Probability Designs

		PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h		
($\times 10^{-3}$)		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	
Normal	NULL	-4.5	29.3	-3.7	33.6	-3.2	30.5	-4.5	29.3	
	Errors	LINDOWN	-0.9	27.0	3.7	35.5	1.8	27.7	-0.7	26.9
	$\tau = 0.2$	EXP	5.8	32.0	1.9	56.8	0.4	39.4	14.1	34.4
	$\sigma = 0.1$	SINE	7.1	30.1	6.1	39.5	3.6	32.8	5.3	30.4
Normal	NULL	-7.7	21.3	-7.7	24.9	-6.6	21.1	-7.6	21.2	
	Errors	LINDOWN	1.1	20.7	3.2	30.6	1.2	20.7	3.5	21.1
	$\tau = 0.1$	EXP	-2.3	20.9	-6.5	53.3	-7.2	30.0	-3.0	20.9
	$\sigma = 0.2$	SINE	5.6	20.9	6.9	36.2	4.0	28.6	4.3	21.1
Log-normal	NULL	-0.5	28.5	-2.0	30.6	-2.1	29.5	-0.3	28.5	
	Errors	LINDOWN	5.4	32.6	5.0	39.0	3.7	34.1	6.0	32.7
	$\tau = 0.2$	EXP	-1.3	28.6	-7.6	62.6	-7.1	36.8	-9.3	30.3
	$\sigma = 0.1$	SINE	3.7	31.2	2.3	43.1	0.1	36.1	1.6	31.0
Log-normal	NULL	3.6	22.8	5.7	28.8	5.7	24.2	3.6	22.7	
	Errors	LINDOWN	6.0	26.8	9.3	37.5	7.5	27.3	2.5	26.0
	$\tau = 0.1$	EXP	0.8	26.3	-2.3	50.8	-3.5	33.1	11.5	29.0
	$\sigma = 0.2$	SINE	3.7	26.9	2.9	37.6	-0.1	30.2	2.2	27.8

Table 3

Variance Estimation and Empirical Coverage Rates of 95% C.I. Using the Model-based, Jackknife and BRR Methods

		Empirical variance	Empirical Bayes Model-based		Jackknife($K = 8$)		BRR		
		Shape	Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%	
Normal	NULL	88	74	92.8	94	96.4	96	94.4	
	Errors	LINDOWN	94	73	89.6	94	94.6	98	94.2
	$\tau = 0.2$	EXP	85	70	91.4	88	94.6	85	93.4
	$\sigma = 0.1$	SINE	83	67	91.6	90	95.8	85	94.4
Normal	NULL	48	45	93.8	48	96.0	49	93.8	
	Errors	LINDOWN	42	45	96.8	51	96.2	51	96.8
	$\tau = 0.1$	EXP	53	54	95.0	61	97.2	59	95.2
	$\sigma = 0.2$	SINE	44	46	95.8	55	96.6	49	96.0
Log-normal	NULL	104	83	91.8	104	94.8	100	93.6	
	Errors	LINDOWN	102	98	93.6	106	95.6	107	95.0
	$\tau = 0.2$	EXP	81	77	93.4	97	96.4	89	94.8
	$\sigma = 0.1$	SINE	92	99	94.8	97	95.2	92	93.4
Log-normal	NULL	93	97	94.2	100	96.2	99	95.2	
	Errors	LINDOWN	104	101	93.6	106	96.0	102	92.8
	$\tau = 0.1$	EXP	84	81	94.6	84	95.2	82	95.0
	$\sigma = 0.2$	SINE	110	96	94.4	98	95.6	92	93.0

Tables 4 and 5 give the empirical variance of the PMM estimator when the non-sampled PSU counts N_h are estimated. They also give the mean estimated variance of this estimator and corresponding coverage rates by the 95% C.I. The confidence intervals are calculated by the usual normal theory intervals based on our point and variance estimators. These two tables show the inference method discussed in section 5 tends to underestimate the true variance of PMM estimator using \hat{N}_h , giving in occasion under-coverage of the population mean. It remains to be studied in the future whether the JRR and BRR methods also yield satisfactory inferences for this method.

For the simulation study using 5% PUMS data, the simple mean has bias = -50.9 and RMSE = 2,600 and the p -spline nonparametric mixed model based method has bias = -41.9 and RMSE = 2,153. Thus both methods have small bias and the model-based estimator has a RMSE 17% less than the RMSE of the simple mean. This improved efficiency is due to the fact that the average household income decreases for as the number of families in the PSUs increases (figure 1). The PMM method exploits this relationship in its predictions.

Table 4

Variance Estimation and Empirical Coverage Rates of 95% C.I. Using *P*-spline and Estimated PSU Counts, Population Simulated with Normal Errors

	$\sigma = 0.1$ and $\tau = 0.2$			$\sigma = 0.2$ and $\tau = 0.1$		
	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate
NULL	90	76	91.8	50	46	93.2
LINDOWN	93	74	90.4	43	46	95.6
EXP	85	72	93.0	55	56	96.2
SINE	110	98	94.8	50	55	97.6

Table 5

Variance Estimation and Empirical Coverage Rates of 95% C.I. Using *P*-spline and Estimated PSU Counts, Population Simulated with Log-normal Errors

	$\sigma = 0.1$ and $\tau = 0.2$			$\sigma = 0.2$ and $\tau = 0.1$		
	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate
NULL	105	84	91.8	95	99	94.8
LINDOWN	103	98	94.4	110	102	94.4
EXP	81	79	94.6	87	83	94.2
SINE	110	150	96.4	91	130	95.8

6. DISCUSSION

Previous parametric model-based inference methods have been criticized mainly for their potentially large design biases when the model is misspecified. In our nonparametric models, the linearity assumption is replaced by a much weaker assumption of a smoothly-varying relationship. As a result, the model-based estimators are more robust, having small biases for a variety of population shapes.

Design information such as inclusion probabilities plays a key role in the model-based inference. Inverse-probability weighted methods imply simple assumptions about the relationship between the outcome variables and the design variables. With the method we propose, the gain in efficiency is realized by applying nonparametric models that relax these assumptions.

Our study has an interesting finding that the model-based estimators can be more efficient than the simple mean for an equal probability design. In other studies, we also find gains in efficiency from *p*-spline nonparametric mixed model in estimating post-stratum means in post-stratified samples.

The empirical Bayes method, the jackknife and BRR methods all give good confidence coverage with confidence intervals that are narrower than those given by the traditional methods. However, we expect the empirical Bayes method to be sensitive to model assumptions on the variance components (e.g., constant within-PSU variances). When the PSU counts are not known for the sample but not for the whole population, model-based estimates of the

unknown counts can still provide sound estimates of the population mean, if the model tracks the true PSU counts precisely enough. The model relating these counts to the auxiliary variable was treated parametrically here, but this could also be specified nonparametrically without much difficulty.

We believe *p*-spline nonparametric mixed models can be applied to more complex designs such as stratified and multi-stage designs. We also believe without much more effort our methods can be generalized for binary or ordinal outcomes.

ACKNOWLEDGEMENTS

This research was supported by grant DMS 0106914 from the National Science Foundation.

REFERENCES

- BRUMBACK, B.A., RUPPERT, D. and WAND, M.P. (1999). Comment to variable selection and function estimation in additive nonparametric regression using data-based prior. *Journal of the American Statistical Association*, 94, 794-797.
- COULL, B.A., SCHWARTZ, J. and WAND, M.P. (2001). Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics*, 2(3), 337-349.
- ELLIOTT, M.R., and LITTLE, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

- FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, B*, 60, 3-21.
- GHOSH, M., and MEEDEN, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.
- HINKLEY, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19, 285-292.
- HOLT, D., and SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society, A*, 142, 33-46.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.
- KALTON, G. (1977). Practical methods for estimating survey sampling errors. *Bulletin of the International Statistical Institute*, 47, 495-514.
- LAZZARONI, L.C., and LITTLE, R.J.A. (1998). Random effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.
- LIN, X., and ZHANG, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society, B*, 61, 381-400.
- LITTLE, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.
- MILLER, R.G. (1974). An unbalanced jackknife. *Annals of Statistics*, 2, 880-891.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag.
- SHAO, J., and WU, C.F.J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models. *Annals of Statistics*, 15, 1563-1579.
- SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.
- U.S. CENSUS (1990). Dept. of Commerce. Census of Population and Housing, [United States]: public use microdata sample; 5- percent sample Computer file]. 3rd release. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 1995. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor], 1996.
- ZHENG, H., and LITTLE, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- ZHENG, H., and LITTLE, R.J.A. (2004). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. To appear in *Journal of Official Statistics*.

A Finite Population Estimation Study with Bayesian Neural Networks

FAMING LIANG and ANTHONY YUNG CHEUNG KUK¹

ABSTRACT

In this article, we study the use of Bayesian neural networks in finite population estimation. We propose estimators for finite population mean and the associated mean squared error. We also propose to use the student t -distribution to model the disturbances in order to accommodate extreme observations that are often present in the data from social sample surveys. Numerical results show that Bayesian neural networks have made a significant improvement in finite population estimation over linear regression based methods.

KEY WORDS: Bayesian model averaging; Bayesian neural network; Evolutionary Monte Carlo; Finite population; Markov Chain Monte Carlo; Prediction.

1. INTRODUCTION

Regression estimation is widely used in sample surveys for incorporating auxiliary population information (Cochran 1977) with the underlying model

$$y_t = \beta_0 + x_{t1}\beta_1 + \dots + x_{tp}\beta_p + \epsilon_t, \quad t = 1, 2, \dots, n, \quad (1)$$

where y_t is the survey variable for the t^{th} element of a population, $\mathbf{x}_t = (x_{t1}, \dots, x_{tp})$ is the vector of auxiliary variables associated with y_t , $\beta_0, \beta_1, \dots, \beta_p$ are the regression coefficients, and ϵ_t is the independent disturbance with zero mean and common variance. Although this model generally performs well, it has several inherent limitations. First, the model is specified linearly and thus can not capture some types of nonlinear relationship, which may be essential in some applications. Second, the least squares estimate, which is widely used for the model (1), may not be reliable in the presence of collinearity among the auxiliary variables. In this case, techniques, such as condition number reduction (Bankier 1990), ridge regression (Bardsley and Chambers 1984), and various variable selection procedures (Silva and Skinner 1997), have to be used to improve the poor prediction performance of the model. Third, in the presence of outliers, the least squares estimate may be severely affected by the outliers.

There are attempts to lessen the dependence of estimators on the linear model (1). Firth and Bennett (1998) identify a sufficient "internal bias calibration" condition under which a model-based estimator is automatically design consistent, regardless of how well the underlying model fits the population. The condition is met by certain estimators based on linear models, certain canonical link generalized linear models and nonparametric regression estimators constructed from them by a particular style of local likelihood fitting.

Bias can also be calibrated externally, if not internally. Chambers, Dorfman and Wehrly (1993) start with a predictor of the population mean based on a heteroscedastic linear model and adjust for its bias using nonparametric regression. Kuk and Welsh (2001) propose a robustified model-based approach whereby a working model is first fitted using robust methods and subsequently the conditional distributions of the residuals given \mathbf{x} are estimated nonparametrically to account for local model departure or outliers in localized regions.

Another way of incorporating auxiliary information into an estimator into an estimator in a design consistent manner is the model-calibrated approach first proposed by Deville and Särndal (1992). The basic idea is to choose weights that satisfy certain calibration equations and are closest to the normal Horvitz-Thompson design weights according to some distance measure. Theberge (1999) applies the calibration technique to estimate population parameters other than the means. More recently, Wu and Sitter (2001) extends the calibration approach to deal with nonlinear as well as generalized linear models by using the fitted values under these working models to set up the calibration equations. The model-calibration approach can be classified as "model-assisted" because while the efficiency of the model-calibrated estimator depends on the validity of the model, consistency does not.

There is certainly a growing trend in the survey literature in using nonlinear and nonparametric regression. Instead of model (1), one considers,

$$y_t = g(\mathbf{x}_t) + \epsilon_t,$$

where the regression function $g(\cdot)$ can be any arbitrary smooth function. Dorfman (1992) estimates g using the Nadaraya-Watson kernel estimator \hat{g} to result in the

¹ Faming Liang, Department of Statistics, Texas A&M University, College Station, TX77843-3143. E-mail: fliang@stat.tamu.edu; Anthony Yung Cheung Kuk, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117543. E-mail: stakuka@nus.edu.sg.

following model-based estimator or predictor of the finite population mean,

$$\hat{y}_K = N^{-1} \left\{ \sum_{t=1}^n y_t + \sum_{t=n+1}^N \hat{g}(x_t) \right\},$$

where it is assumed without loss of generality that the sample consists of the first n elements of the population. Kuk (1993) makes use of kernel method to estimate the conditional distribution of y given x as a way of incorporating auxiliary information in the estimation of the finite population distribution of y . For the case of scalar x , Breidt and Opsomer (2000) estimates g using local polynomial regression with design weights incorporated to account for the sampling design used and propose a generalized difference estimator,

$$\hat{y}_{LP} = N^{-1} \left\{ \sum_{t=1}^n \frac{y_t - \hat{g}(x_t)}{\pi_t} + \sum_{t=1}^N \hat{g}(x_t) \right\} = N^{-1} \left\{ \sum_{t=1}^n w_t y_t \right\},$$

where π_t is the sample inclusion probability. It can be shown that the weights w_t are calibrated to match the totals of x up to the q^{th} order, where q is the order of the local polynomial. As a consequence, \hat{y}_{LP} is exactly model-unbiased if the true regression function is a polynomial of degree q or less. Breidt and Opsomer (2000) also show that \hat{y}_{LP} is asymptotically design-unbiased and consistent under mild conditions. For more discussions on nonlinear and nonparametric methods, see Valliant, Dorfman and Royall (2000) (chapter 11).

In this paper, another nonlinear regression method, Bayesian neural network (BNN), is applied to the problem. BNN has an important advantage of being able to handle multivariate auxiliary variables and model selection with ease, which is not the case for many other nonlinear and nonparametric techniques. BNNs were first introduced by Buntine and Weigend (1991) and MacKay (1992), and were further developed by Neal (1996), Müller and Insua (1998), Marrs (1998), Holmes and Mallick (1998), and Liang and Wong (2001). But the BNN proposed in this paper is different from those cited above in one important respect: A prior is put on each network connection, instead of only on the number of hidden units as done in the literature. This allows us to treat the selection of network structure and the selection of input variables (auxiliary variables) uniformly. The network is trained by sampling from the joint posterior of the network structure and connection weights. The sampled network has often a sparse structure, which effectively prevents the data from being overfitted. A heavy tail distribution, such as the student t -distribution, is proposed to model the disturbances of the data with outliers. Numerical results show that BNN models have offered a significant improvement over the linear regression based models in finite population estimation.

The remaining part of this article is organized as follows. In section 2, we describe the BNN models and the associated estimators for finite populations. In section 3, we present our numerical results for one finite population example with two choices of auxiliary variables and comparisons with various linear regression based models. In section 4, we present our numerical results for another finite population example demonstrate how a cross-validation procedure can be applied to determine the parameter setting for BNN models. In section 5, we conclude the paper with a brief discussion.

2. FINITE POPULATION ESTIMATION WITH BAYESIAN NEURAL NETWORKS

2.1 Bayesian Neural Network Models

Suppose we have data pairs $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, which were generated from the relationship

$$y_t = g(x_t) + \epsilon_t, \quad (2)$$

where $y_t \in R^1$, $x_t = (x_{t1}, \dots, x_{tp}) \in R^p$, $g(\cdot)$ is the true regression function of unknown form, and $\epsilon_t/\sigma \sim t(\nu)$ with $\nu > 2$ being a known degree of freedom of the t -distribution. Here $g(\cdot)$ may be highly nonlinear, and σ is an unknown scale parameter. We use the student t -distribution, instead of the Gaussian distribution as usual, to model the disturbances in order to accommodate extreme observations that are often present in the data from social sample surveys.

Before describing our BNN model, we first give a brief description for feed-forward neural networks. Figure 1 illustrates a one-hidden layer feed-forward neural network. It consists of four types of units, bias units, input units, hidden units, and output units. The unit to which the input features are presented is referred to as the input unit. The bias unit is a special type of input units with a constant input, say, 1. The unit where the network output is formed is referred to as the output unit. The hidden unit is so called because its input and output are only used for internal connections and are unavailable to the outside world. In a feed-forward neural network, each hidden unit independently processes the values fed to it by the units in the preceding layer and then presents its output to the units in the next layer for further processing. It has been shown by several authors (Cybenko 1989; Funahashi 1989; Hornik, Stinchcombe and White 1989) that neural networks are universal approximators in that a one-hidden layer feed-forward neural network with linear output units can approximate any continuous functions arbitrarily well on compact sets by increasing the number of hidden units. To survey regression, this is an important advantage of neural network models over other regression models. In the survey regression literature, whether model-assisted or model-based,

there is usually considerable attention paid to the consequences of model misspecification. The neural network model avoids this consideration partially due to its specific property of universal approximation. In section 2.2.1, we show that as the sample size is large, the unknown regression function $g(\cdot)$ in (2) can be well approximated by BNN models, regardless of the true function form of $g(\cdot)$. Essentially, BNN falls into the class of data-driven methods.



Figure 1. A fully connected one hidden layer feed-forward neural network with 4 input units, 3 hidden units and 1 output unit. The arrows indicate the direction of data feeding.

In our BNN model, the function $g(\cdot)$ in model (2) is approximated by a function of the form

$$\hat{g}(\mathbf{x}_t, \mathbf{a}, \mathbf{\beta}, \boldsymbol{\gamma}) = \alpha_0 I_{\alpha_0} + \sum_{i=1}^p x_{ti} \alpha_i I_{\alpha_i} + \sum_{j=1}^M \beta_j I_{\beta_j} \psi \left(\sum_{i=1}^p x_{ti} \gamma_{ji} I_{\gamma_{ji}} + \gamma_{j0} I_{\gamma_{j0}} \right), \quad (3)$$

where I_{ζ} is an indicator function which indicates the effectiveness of the connection ζ ; M denotes the maximum number of hidden units which is specified by users; α_0 denotes the bias term of the output unit; $\alpha_1, \dots, \alpha_p$ denote the weights on the connections from the input units to the output unit; β_1, \dots, β_M denote the weights on the connections from hidden units to the output unit; γ_{j0} denotes the bias term of the j^{th} hidden unit; $\gamma_{j1}, \dots, \gamma_{jp}$ denote the weights on the connections from the input units to the j^{th} hidden unit; and $\psi(\cdot)$ denotes the activation function. Sigmoid and hyperbolic tangent functions are two popular choices for the activation function. We set $\psi(z) = \tanh(z)$ for all examples of this paper.

Let Λ be the vector consisting of all indicators of model (3). Note that Λ specifies the structure of the corresponding network. Let $\mathbf{a} = (\alpha_0, \alpha_1, \dots, \alpha_p)$, $\mathbf{\beta} = (\beta_1, \dots, \beta_M)$, $\boldsymbol{\gamma}_i = (\gamma_{i0}, \dots, \gamma_{ip})$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_M)$, and $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{\beta}, \boldsymbol{\gamma}, \sigma^2)$, where $\alpha_\Lambda, \beta_\Lambda$ and γ_Λ denote the non-zero subsets of $\mathbf{a}, \mathbf{\beta}$ and $\boldsymbol{\gamma}$, respectively. Thus, the model (3) is completely

specified by the tuple $(\boldsymbol{\theta}, \Lambda)$. For simplicity, in the following we will use $\boldsymbol{\theta}_\Lambda$ to denote a BNN model and use $\hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_\Lambda)$ to re-denote the function $\hat{g}(\mathbf{x}_t, \mathbf{a}, \mathbf{\beta}, \boldsymbol{\gamma})$. Also, we let $\boldsymbol{\theta}_\Lambda = (\boldsymbol{\theta}, \Lambda)$, and use $\boldsymbol{\theta}_\Lambda$ and $(\boldsymbol{\theta}, \Lambda)$ exchangeably. To conduct a Bayesian analysis for model (3), we have the following prior distributions: $\alpha_i \sim N(0, \sigma_\alpha^2)$ for $\alpha_i \in \alpha_\Lambda$; $\beta_j \sim N(0, \sigma_\beta^2)$ for $\beta_j \in \beta_\Lambda$; $\gamma_{ji} \sim N(0, \sigma_\gamma^2)$ for $\gamma_{ji} \in \gamma_\Lambda$; and $f(\sigma^2) \sim 1/\sigma^2$. The total number of effective connections in Λ is $m = \sum_{i=0}^p I_{\alpha_i} + \sum_{j=1}^M I_{\beta_j} \delta(\sum_{i=0}^p I_{\gamma_{ji}}) + \sum_{j=1}^M \sum_{i=0}^p I_{\beta_j} I_{\gamma_{ji}}$, where $\delta(z) = 1$ if $z > 0$ and 0 otherwise. The model Λ is subject to a prior probability that is proportional to the mass put on m by a truncated Poisson (λ) with rate λ ,

$$P(\Lambda) = \begin{cases} \frac{1}{Z} \frac{\lambda^m}{m!}, & m = 3, 4, \dots, U \\ 0, & \text{otherwise} \end{cases}$$

where $U = (M+1)(p+1) + M$ is the number of connections of the full model in which all $I_\zeta = 1$; and $Z = \sum_{\Lambda \in \Omega} \lambda^m / m!$. Here we let Ω denote the set of all possible models with $3 \leq m \leq U$. We set the minimum number of m to three based on our views: neural networks are usually used for complex problems, and three has been a small enough number as a limiting network size. In these prior distributions, $\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2$ and λ are hyper-parameters to be specified by users (discussed below). Furthermore, we assume that these prior distributions are independent *a priori*. Thus, we have the following log-posterior (up to an additive constant),

$$\begin{aligned} \log \pi(\boldsymbol{\theta}_\Lambda | D) = & \text{Constant} - \left(\frac{n}{2} + 1 \right) \log \sigma^2 - \frac{\nu + 1}{2} \sum_{t=1}^n \log \left(1 + \frac{(y_t - \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_\Lambda))^2}{\nu \sigma^2} \right) \\ & - \frac{1}{2} \sum_{i=0}^p I_{\alpha_i} \left(\log \sigma_\alpha^2 + \frac{\alpha_i^2}{\sigma_\alpha^2} \right) - \frac{1}{2} \sum_{j=1}^M I_{\beta_j} \left(\delta \left(\sum_{i=0}^p I_{\gamma_{ji}} \right) \right. \\ & \quad \left. \left(\log \sigma_\beta^2 + \frac{\beta_j^2}{\sigma_\beta^2} \right) \right) \\ & - \frac{1}{2} \sum_{j=1}^M \sum_{i=0}^p I_{\beta_j} I_{\gamma_{ji}} \left(\log \sigma_\gamma^2 + \frac{\gamma_{ji}^2}{\sigma_\gamma^2} \right) - \frac{m}{2} \log(2\pi) \\ & + m \log \lambda - \log(m!). \end{aligned} \quad (4)$$

Our BNN model is different from other BNN models existing in the literature in two important respects. First, the input variables of our BNN model are selected automatically by sampling from the joint posterior of the network structure and weights. Second, the structure of our BNN model is usually sparse and its performance less depends on

the initial specification for the input patterns and the number of hidden units. The sparse is in the sense that only a small number of connections are active in the network. So our BNN model avoids the problem of overfitting in a more natural way.

For data preparation and hyperparameter setting, we have the following suggestions. To avoid some weights that are trained to be extremely large or small (in absolute value) to accommodate different scales of input and output variables, we suggest that all input and output variables be normalized before feeding to the networks. In all examples of this article, the data is normalized by $(y_i - \bar{y})/S_y$, where \bar{y} and S_y denote the mean and standard deviation of the training data, respectively. Based on the belief that a network with a large weight variation usually has a poor generalization performance, we suggest that $\sigma_\alpha^2, \sigma_\beta^2$ and σ_γ^2 are chosen for moderate values to penalize a large weight variation. For example, we set $\sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\gamma^2 = 5$ for all examples of this article. The setting should also be fine for the other problems. The value of λ reflects our belief on the network size needed for the data under consideration. Here we follow the suggestion of Weigend, Huberman and Rumelhart (1990) to choose λ such that the number of connection weights is about one tenth of the size of the training sample. In one simulation, we assessed the influence of λ on BNN model size and prediction ability. The numerical results suggest that the prediction ability of BNN models is rather robust to the variation of λ , although the BNN model size increases slowly as λ increases.

To sample from the posterior (4), a Monte Carlo algorithm, so called the reversible jump evolutionary Monte Carlo (RJEMC) algorithm, is developed. This algorithm extends the evolutionary Monte Carlo algorithm (Liang and Wong 2001) to sample from a variable dimensional space by incorporating some reversible jump moves proposed in Green (1995). For details of the algorithm, please refer to the support documents and software for the paper. They are available at <http://www.stat.tamu.edu/~fliang>.

2.2 Finite Population Estimation with Bayesian Neural Networks

2.2.1 Bayesian Model Averaging

In this subsection, we review some basic results of Bayesian model averaging and show one theorem for BNN models, which form the theoretical basis for the estimators described in section 2.2.2. Suppose that we are interested in estimating the quantity $p(\theta_\Lambda)$, which is a function of both Λ and θ . The Bayesian estimator of $p(\theta_\Lambda)$ can be written as

$$E_\pi p(\theta_\Lambda) = \sum_{k=0}^K P(\Lambda_k|D) \int p(\theta_k, \Lambda_k) \pi(\theta_k|\Lambda_k, D) d\theta_k, \quad (5)$$

where K denotes the total number of models under consideration, θ_k denotes the parameters associated with model

Λ_k , and $\pi(\theta_k|\Lambda_k, D)$ denotes the posterior density of θ_k conditional on model Λ_k . Madigan and Raftery (1994) argued for this estimator that Bayesian model averaging (averaging over all the models in this fashion) accounts for the model uncertainty, and provides better predictive ability, as measured by the logarithmic scoring rule, than using any single model Λ_k . See Hoeting, Madigan, Raftery and Volinsky (1999) for a tutorial on Bayesian model averaging.

Suppose that samples $(\theta_1, \Lambda_1), \dots, (\theta_M, \Lambda_M)$ have been drawn from the posterior distribution $\pi(\theta_\Lambda|D)$ by a MCMC algorithm, then $p(\theta_\Lambda)$ can be estimated by

$$\hat{p}(\theta_\Lambda) = \frac{1}{M} \sum_{i=1}^M p(\theta_{\Lambda_i}), \quad (6)$$

where $\theta_\Lambda = (\theta_1, \Lambda_1)$. Applying the standard Markov chain theory (Tierney 1994; Roberts and Casella 1999), under regularity conditions we have the following results. If $E_\pi |p(\theta_\Lambda)| < \infty$, then

$$\frac{1}{M} \sum_{i=1}^M p(\theta_{\Lambda_i}) \rightarrow E_\pi p(\theta_\Lambda), \quad \text{a.s.}, \quad (7)$$

as $M \rightarrow \infty$. Furthermore, if $E_\pi |p(\theta_\Lambda)|^{2+\delta} < \infty$ for some $\delta > 0$, then

$$M^{1/2} \left\{ \frac{1}{M} \sum_{i=1}^M p(\theta_{\Lambda_i}) - E_\pi p(\theta_\Lambda) \right\} \rightarrow N(0, \tau^2), \quad (8)$$

for some positive constant τ^2 as $M \rightarrow \infty$, and the convergence is in distribution.

Similar to (7) and (8), we have the following theorem for BNN models, of which proof is presented in Appendix.

Theorem 2.1 Let $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ denote a simple random sample drawn from a population which can be modeled by model (2). Let $(\theta_1, \Lambda_1), \dots, (\theta_M, \Lambda_M)$ denote the sample drawn from the posterior distribution $\pi(\theta_\Lambda|D)$, given in (4), by a MCMC method. Then, for any x_0 drawn from the same distribution with the observations D , we have

$$(a) \quad E_\pi | \hat{g}(x_0, \theta_\Lambda) |^{2+\delta} < \infty, \quad (9)$$

for some $\delta > 0$, as $n \rightarrow \infty$.

$$(b) \quad \frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) \rightarrow g(x_0), \quad \text{a.s.}, \quad (10)$$

as $n \rightarrow \infty$ and $M \rightarrow \infty$.

$$(c) \quad M^{1/2} \left[\frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) - g(x_0) \right] \rightarrow N(0, \tau_*^2), \quad (11)$$

for some positive constant τ_*^2 as $n \rightarrow \infty$ and $M \rightarrow \infty$, and the convergence is in distribution.

To show some properties of moments of $1/M \sum_{i=1}^M \hat{g}(\mathbf{x}_0, \boldsymbol{\theta}_{\Lambda_i})$, we need the following theorem (Billingsley 1986, page 348, Corollary),

Theorem 2.2 *Let r be a positive integer. If $X_M \rightarrow X$ in distribution and $\sup_m E|X_m|^{r+\delta} < \infty$, where $\delta > 0$, then $E|X|^r < \infty$ and $EX_m^r \rightarrow EX^r$.*

Following from (9), (11) and Theorem 2.2, we know

$$ME \left[\frac{1}{M} \sum_{i=1}^M \hat{g}(\mathbf{x}_0, \boldsymbol{\theta}_{\Lambda_i}) - g(\mathbf{x}_0) \right]^2 \rightarrow \tau_*^2, \quad (12)$$

as $n \rightarrow \infty$ and $M \rightarrow \infty$. It implies that

$$E \left[\frac{1}{M} \sum_{i=1}^M \hat{g}(\mathbf{x}_0, \boldsymbol{\theta}_{\Lambda_i}) - g(\mathbf{x}_0) \right]^2 = \frac{\tau_*^2}{M} + o\left(\frac{1}{M}\right) \quad (13)$$

holds as n and M are both large.

Note we have shown that (11) and (13) hold as the sample size $n \rightarrow \infty$. In the context of finite population, especially for a small finite population, a more precise expression for (11) and (13) would be

$$M^{1/2} \left[\frac{1}{M} \sum_{i=1}^M \hat{g}(\mathbf{x}_0, \boldsymbol{\theta}_{\Lambda_i}) - E(y_0 | D, \mathbf{x}_0) \right] \rightarrow N(0, \tau_*^2), \quad (14)$$

and

$$E \left[\frac{1}{M} \sum_{i=1}^M \hat{g}(\mathbf{x}_0, \boldsymbol{\theta}_{\Lambda_i}) - E(y_0 | D, \mathbf{x}_0) \right]^2 = \frac{\tau_*^2}{M} + o\left(\frac{1}{M}\right), \quad (15)$$

where $E(y_0 | D, \mathbf{x}_0)$ denotes the prediction of y_0 which is the survey variable corresponding to \mathbf{x}_0 . The equations (14) and (15) take into accounts the possible bias of the sample D . In the case that the population constitutes many exact copies of the sample D , $E(y_0 | D, \mathbf{x}_0) = g(\mathbf{x}_0)$ holds, and equations (14) and (15) are reduced to (11) and (13), respectively.

2.2.2 BMA Estimators in Finite Populations

Consider a finite population of N distinguishable elements. Associated with the i^{th} elements are the survey variable y_i and the auxiliary variables \mathbf{x}_i . The values $\mathbf{x}_1, \dots, \mathbf{x}_N$ are known for the entire population, while y_i is known only if the i^{th} unit is selected in the sample. Suppose a simple random sample $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ has been drawn from the finite population, a BNN model has been built for the sample, and $(\boldsymbol{\theta}_1, \Lambda_1), \dots, (\boldsymbol{\theta}_M, \Lambda_M)$ have been drawn from the posterior distribution of the BNN model, the BMA estimator for the mean of the finite population is

$$\bar{y}_{\text{BNN}} = f \bar{y} + \frac{1}{MN} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_{\Lambda_i}),$$

where \bar{y} is the sample mean of y_1, \dots, y_n , and $f = n/N$ is the sample fraction. About this estimator, we have the

following comments. First, \bar{y}_{BNN} is a model-based estimator, so that all the inference is with respect to the model for the y_i 's, not the survey design. As long as the model holds, the BNN estimator will have the mean squared error properties described below for any ignorable sampling design. Second, this estimator is identical to that proposed in Dorfman (1992), except that the BNN is replaced by a kernel-based regression. Third, this estimator can be used to estimate the mean of a finite population as long as each of the unsampled elements has the same distribution as the sample D .

The accuracy of an estimate can be measured by its mean squared error $E(\bar{y}_{\text{BNN}} - \bar{Y})^2$, where \bar{Y} denotes the true population mean. To estimate $E(\bar{y}_{\text{BNN}} - \bar{Y})^2$, we first consider

$$\begin{aligned} & E[(\bar{y}_{\text{BNN}} - \bar{Y})^2 | D, \mathbf{X}_{n+1}^N] \\ &= E \left[\left\{ \frac{1}{MN} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_{\Lambda_i}) - \frac{1}{N} \sum_{t=n+1}^N (g(\mathbf{x}_t) + \epsilon_t) \right\}^2 \middle| D, \mathbf{X}_{n+1}^N \right] \\ &= \frac{(N-n)^2}{N^2} E \left[\left\{ \frac{1}{M(N-n)} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_{\Lambda_i}) - \frac{1}{N-n} \sum_{t=n+1}^N g(\mathbf{x}_t) \right\}^2 \middle| D, \mathbf{X}_{n+1}^N \right] \\ &\quad + \frac{N-n}{N^2} \text{var}(\epsilon_t) \\ &= \frac{(N-n)^2}{N^2} E \left[\left\{ \frac{1}{M(N-n)} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_{\Lambda_i}) - E(\bar{y}_u | D, \mathbf{X}_{n+1}^N) + E(\bar{y}_u | D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u) \right\}^2 \middle| D, \mathbf{X}_{n+1}^N \right] \\ &\quad + \frac{N-n}{N^2} \text{var}(\epsilon_t) \\ &\approx \frac{\tau_D^2}{M} + (1-f)^2 \{ E(\bar{y}_u | D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u) \}^2 \\ &\quad + \frac{1-f}{N} \text{var}(\epsilon_t), \end{aligned} \quad (16)$$

where $\mathbf{X}_{n+1}^N = (\mathbf{x}_{n+1}, \dots, \mathbf{x}_N)$ denotes the set of auxiliary vectors of the unsampled elements; \bar{y}_u denotes the averaged survey value of the unsampled elements, and

$$E(\bar{y}_u) = \frac{1}{N-n} \sum_{t=n+1}^N g(\mathbf{x}_t).$$

The last approximation of (16) follows from (15), that is, as M is large,

$$E\left\{\frac{1}{MN} \sum_{i=1}^M \sum_{t=n+1}^M \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_{\Lambda_t}) - (1-f)E(\bar{y}_u | D, \mathbf{X}_{n+1}^N)\right\}^2 \approx \frac{\tau_D^2}{M},$$

for some positive constant τ_D^2 . The term $E(\bar{y}_u | D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u)$ is the prediction bias due to the randomness or sampling bias of D . Following from (16), we have

$$E(\bar{y}_{\text{BNN}} - \bar{Y})^2 \approx \frac{E\tau_D^2}{M} + (1-f)^2$$

$$E\left\{E(\bar{y}_u | D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u)\right\}^2 + \frac{1-f}{N} \text{var}(\epsilon_t). \quad (17)$$

The quantity τ_D^2 can be estimated by the batch means method (Roberts 1996) as follows. Run the Markov chain for $M = rs$ iterations, where s is the batch size and is assumed sufficiently large such that

$$\bar{y}_{\text{BNN},k} = f\bar{y} + \frac{1}{sN} \sum_{i=(k-1)s+1}^{ks} \sum_{t=n+1}^N \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_{\Lambda_t}),$$

is approximately independently $N(f\bar{y} + (1-f)E(\bar{y}_u | D, \mathbf{X}_{n+1}^N), \tau_D^2/s)$. Therefore τ_D^2 can be approximated by

$$\hat{\tau}_D^2 = \frac{s}{r-1} \sum_{k=1}^r (\bar{y}_{\text{BNN},k} - \bar{y}_{\text{BNN}})^2, \quad (18)$$

which can be substituted into (17) in lieu of $E\tau_D^2$. Under the assumption $\epsilon_t/\sigma \sim t(\nu)$, the BMA estimator $\text{var}(\epsilon_t)$ is

$$\hat{\text{var}}(\epsilon_t) = \frac{\nu}{\nu-2} \frac{1}{M} \sum_{i=1}^M \hat{\sigma}_i^2. \quad (19)$$

Under the assumption that the population is made up of exact copies of the training data, we have $E(\bar{y}_u | D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u) \approx \bar{\hat{y}} - \bar{y}$, where $\bar{\hat{y}}$ denotes the fitted sample mean, and

$$E(\bar{\hat{y}} - \bar{y})^2 = E\left\{\frac{1}{n} \sum_{t=1}^n \hat{\epsilon}_t\right\}^2 = \frac{1}{N} \text{var}(\hat{\epsilon}_t), \quad (20)$$

where $\hat{\epsilon}_t = \sum_{i=1}^M \hat{g}(\mathbf{x}_t, \boldsymbol{\theta}_{\Lambda_t})/M - y_t$ is the residual of the t^{th} element of D , and $\hat{\epsilon}_t$'s are assumed to be iid and $E(\hat{\epsilon}_t) = 0$. Under the true model, we have $\text{var}(\hat{\epsilon}_t) \approx \text{var}(\epsilon_t)$. Hence, we suggest $E\{E(\bar{y}_u | D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u)\}^2$ be estimated by

$$\hat{\text{Bias}}^2 = \frac{1}{n} \hat{\text{var}}(\epsilon_t). \quad (21)$$

In summary, $E(\bar{y}_{\text{BNN}} - \bar{Y})^2$ can be estimated by

$$\hat{E}(\bar{y}_{\text{BNN}} - \bar{Y})^2 = \frac{\hat{\tau}_D^2}{M} + (1-f)^2 \hat{\text{Bias}}^2$$

$$+ \frac{1-f}{N} \hat{\text{var}}(\epsilon_t) = \frac{\hat{\tau}_D^2}{M} + \frac{1-f}{n} \hat{\text{var}}(\epsilon_t). \quad (22)$$

As $M \rightarrow \infty$ we have

$$\hat{E}(\bar{y}_{\text{BNN}} - \bar{Y})^2 = \frac{1-f}{n} \hat{\text{var}}(\epsilon_t). \quad (23)$$

We note that this estimate is identical in form to that given by Cochran (1977) for the linear regression estimator.

3. FIRST SIMULATION STUDY

3.1 The Data

Our simulation population comprises 426 records for heads of household surveyed using the sample (long) questionnaire during the 1988 Test Population Census of Limeira, in São Paulo state, Brasil. This test was carried out as a pilot survey during the preparation for the 1991 Brazilian Population Census. For a detailed description for the test census, see Silva and Skinner (1997). We followed Silva and Skinner (1997) to consider the total monthly income as the main survey variable (y) together with 11 potential auxiliary variables, namely,

x_1	indicator of sex of head of household equal male;
x_2	indicator of age of head of household less than or equal to 35;
x_3	indicator of age of head of household greater than 35 and less than or equal to 55;
x_4	total number of rooms in household;
x_5	total number of bathrooms in household;
x_6	indicator of ownership of household;
x_7	indicator that household type is house;
x_8	indicator of ownership of at least one car in household;
x_9	indicator of ownership of color TV in household;
x_{10}	years of study of head of household;
x_{11}	proxy of total monthly income of head of household.

Figure 2, the scatter plots of y versus the 11 auxiliary variables, shows that a linear regression model is not appropriate for the data. Although y and x_{11} are strongly linearly correlated, the scatter plots of y versus some other auxiliary variables, say x_4, x_5 and x_{10} , suggest that their relationships can not be well modeled by a linear regression. In addition, if the data is modeled by a linear regression, the outlier, the 53th element, may have a high influence on fitting and prediction of the model. More precisely, if the element is included in the training data, the fitted response curve will have a up-drift comparing to the true curve and as a result the finite population mean will be overestimated; if

the element is not included in the training data, prediction will proceed as though there were not outliers and as a result the finite population mean will be underestimated. The presence of the strong influence element also mounts a great challenge on BNN models and other data analysis strategies.

We followed Silva and Skinner (1997) to construct two alternative sets of auxiliary variables for simulations. The first set contains x_1, \dots, x_4 and x_{11} , which includes the proxy variable x_{11} and has a reasonable explanatory power in predicting y . The second set contains x_1, \dots, x_{10} , which has a weaker explanatory power than the first one due to the exclusion of x_{11} . So these two sets illustrate the predictive performances of BNN models with strong and weak auxiliary variables, respectively. As in Silva and Skinner (1997), 1,000 sample replicates of size 100 from this simulation population are selected by simple random sampling without replacement. The following computation were performed on the 1,000 replicates.

For each replicate, say k , it was analyzed by BNN models and various linear regression based strategies (reviewed below). For any strategy, the population mean estimate and its estimated mean squared error for the replicate k are denoted by $\bar{y}(k)$ and $V(\bar{y}(k))$, respectively.

The computational results were summarized by computing the mean (MEAN), bias (BIAS), mean square error (MSE) and average of mean squared error estimates (AVMSE) from the set of the 1,000 replicates, given respectively by

$$\text{MEAN} = \sum_{k=1}^S \bar{y}(k) / S;$$

$$\text{BIAS} = \text{MEAN} - \bar{Y};$$

$$\text{MSE} = \sum_{k=1}^S [\bar{y}(k) - \bar{Y}]^2 / S;$$

$$\text{AVMSE} = \sum_{k=1}^S V(\bar{y}(k)) / S,$$

where S is the total number of sample replicates under consideration, and $\bar{Y} = 194.34$ for the simulation population. Empirical coverage rates for 95% confidence intervals based on asymptotic normal theory were also computed for each strategy and these rates, expressed as percentages, are presented in the last columns of Tables 1 and 3.

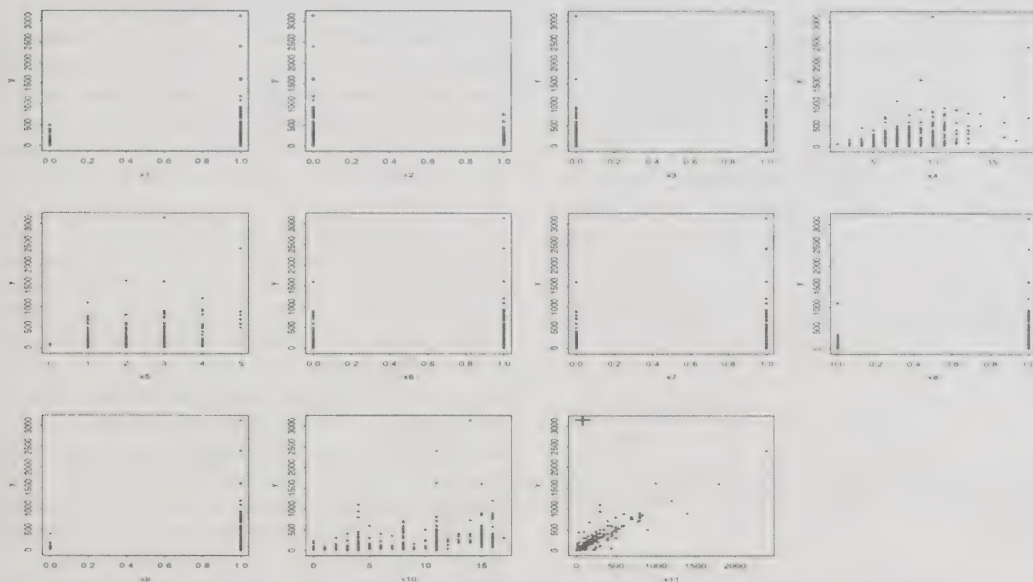


Figure 2. Scatter plots of the response variable y versus the auxiliary variables. In the plot of y versus x_{11} the “+” represents the 53rd element of the population.

3.2 Review of the Linear Regression Based Strategies

The linear regression based strategies that have been considered by Silva and Skinner (1997) are listed as follows.

SM)	Sample mean estimator, with no auxiliary variables (\bar{y}_r, V_s).
Fs)	Forward selection of auxiliary variables with (\bar{y}_r, V_s).
Fd)	Forward selection of auxiliary variables with (\bar{y}_r, V_d).
Fg)	Forward selection of auxiliary variables with (\bar{y}_r, V_g).
Bs)	Best subset selection from all subsets of auxiliary variables with (\bar{y}_r, V_s).
Bd)	Best subset selection from all subsets of auxiliary variables with (\bar{y}_r, V_d).
Bg)	Best subset selection from all subsets of auxiliary variables with (\bar{y}_r, V_g).
FI)	Fixed subset of auxiliary variable with (\bar{y}_r, V_s).
SS)	Saturated subset of auxiliary variable with (\bar{y}_r, V_s).
FR)	Forward subset selection using SAS PROC REG, with (\bar{y}_r, V_s).
CN)	Condition number reduction subset selection procedure with (\bar{y}_r, V_s).
RI)	Ridge regression estimator proposed by Dunstan and Chambers (1986).

To facilitate the description for the above strategies, we define the following notations. Let $U = \{1, \dots, N\}$ denote a finite population of N distinguishable elements, $D \subset U$ denote a sample replicate of n elements drawn from U by simple random sampling without replacement, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ be the vector of auxiliary variables associated with the i^{th} element, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$. Let $\bar{\mathbf{X}} = N^{-1} \sum_{i \in U} \mathbf{x}_i$ be the vector of population means, $\bar{\mathbf{x}} = n^{-1} \sum_{i \in D} \mathbf{x}_i$ be the vector of sample means, $\bar{y} = n^{-1} \sum_{i \in D} y_i$ be the sample mean of the response variable, $\hat{S}_x = n^{-1} \sum_{i \in D} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, $\hat{S}_{xy} = n^{-1} \sum_{i \in D} (\mathbf{x}_i - \bar{\mathbf{x}})(y_i - \bar{y})$, $g_i = 1 + (\bar{\mathbf{X}} - \bar{\mathbf{x}})' \hat{S}_x^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ the so-called g_i -weights (Särndal, Swensson and Wretman 1989), and $\hat{\boldsymbol{\beta}} = \hat{S}_x^{-1} \hat{S}_{xy}$ the least squares estimator of $\boldsymbol{\beta}$. The regression estimator of \bar{Y} is

$$\bar{y}_r = \bar{y} + (\bar{\mathbf{X}} - \bar{\mathbf{x}})' \hat{\boldsymbol{\beta}}.$$

The V_s, V_d is and V_g are three estimators of the mean squared error of \bar{y}_r . The V_s is given by Cochran (1977, page 195),

$$V_s = \frac{1-f}{n(n-p-1)} \sum_{i \in D} \hat{\epsilon}_i^2,$$

where $\hat{\epsilon}_i = (y_i - \bar{y}) - (\mathbf{x}_i - \bar{\mathbf{x}})' \hat{\boldsymbol{\beta}}$ and $f = n/N$ is the sample fraction. The V_d is generalized (from $p=1$ to $p>1$) from one estimator studied by Deng and Wu (1987) and it is expected to have a smaller bias than V_s (Silva 1996),

$$V_d = \frac{1-f}{n(n-1)} \sum_{i \in D} \alpha_i \hat{\epsilon}_i^2,$$

where

$$\alpha_i = (g_i^2 - 2g_i f + f) / \left\{ (1-f) \left[1 - (\mathbf{x}_i - \bar{\mathbf{x}})' \hat{S}_x^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) / (n-1) \right] \right\}.$$

The V_g is modified from one estimator given by Särndal *et al.* (1989), and it has a similar performance to V_d ,

$$V_g = \frac{1-f}{n(n-p-1)} \sum_{i \in D} g_i^2 \hat{\epsilon}_i^2.$$

The best subset selection strategy (Bs, Bd and Bg) is to choose one subset which has the smallest mean squared error estimate among all 2^p possible subsets. The forward selection strategy (Fs, Fd and Fg) starts with the sample mean as an estimator, then adds the variable which minimizes the mean squared error estimate, and the procedure is repeated until the mean squared error estimate starts to increase. Refer to Silva and Skinner (1997) for details of the implementations of the strategies CN and RI.

3.3 Illustration on One Sample Replicate

To understand the behavior of \bar{y}_{BNN} in presence of outliers and the role played by ν in robust inference, we focus on one particular sample. The training data comprises the first 100 elements of the population, and the auxiliary variables include x_1, \dots, x_4 and x_{11} as the first explanatory set. Note that the 53th element has been included in the training data.

For BNN models, we set $\lambda=5$ and $M=8$ which produces 62 connections for the full BNN model, and tried $\nu=25, 50, 100, 200$ and $+\infty$, where $\nu=+\infty$ is equivalent to the assumption $\epsilon_i \sim N(0, \sigma^2)$. For each setting, RJEMC was run as follows: the network connections were first set to some random numbers drawn from $N(0, 0.01)$, and then were updated for 1,000 iterations in the parameter space of the full model, *i.e.*, all indicator variables are set to 1 in those iterations. After the initialization process, 4,000 iterations of RJEMC were run, and 800 samples were collected from these iterations at the lowest temperature level with an equal time space. The convergence of RJEMC can be diagnosed using the Gelman-Rubin statistic \hat{R} (Gelman and Rubin 1992) based on multiple independent runs. Figure 3 shows \hat{R} values computed from 10 independent runs. For each sample replicate of the simulation population, RJEMC converges ($\hat{R} < 1.1$) very fast, usually within the first 500 iterations (100 BNN samples). We discarded the first 200 samples for the burn-in process, and used the remaining 600 samples for the further inference.

For comparison, the linear regression model (1) was also applied to this sample replicate.

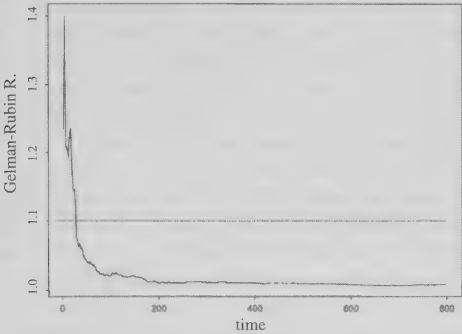


Figure 3. Gelman-Rubin statistic \hat{R} . The curve was computed based on 10 independent runs of RJEMC. The random errors are assumed to be distributed according to $t(100)$.

Figure 4 shows the original data together with the fitted and predicted values produced by various models. The BNN

results were all obtained in one run of RJEMC. It can be seen that the linear regression model is not appropriate for this population as some fitted and predicted values produced by the model are negative for this sample replicate. Also, the fitted response curve (the solid curve in Figure 4(a) and 4(b)) is strongly influenced by the 53th element and lies above almost two-thirds of the data points. A similar phenomenon occurs for the prediction of unsampled values, see Figure 4(c) and 4(d). As a result, the population mean is overestimated (Figure 5). Comparing to that of the linear regression model, the results of the BNN models are less affected by the 53th element, especially for those computed with small values of ν . Figure 5 shows that as ν decreases, the estimated population mean by BNN models gets closer and closer to the true value, and the estimated 95% confidence interval of the population mean becomes narrower and narrower. It indicates that the influence of the 53th element on these estimates becomes weaker and weaker as ν decreases. This is not surprising as the use of a heavily tailed error distribution is known to make the inference more robust.

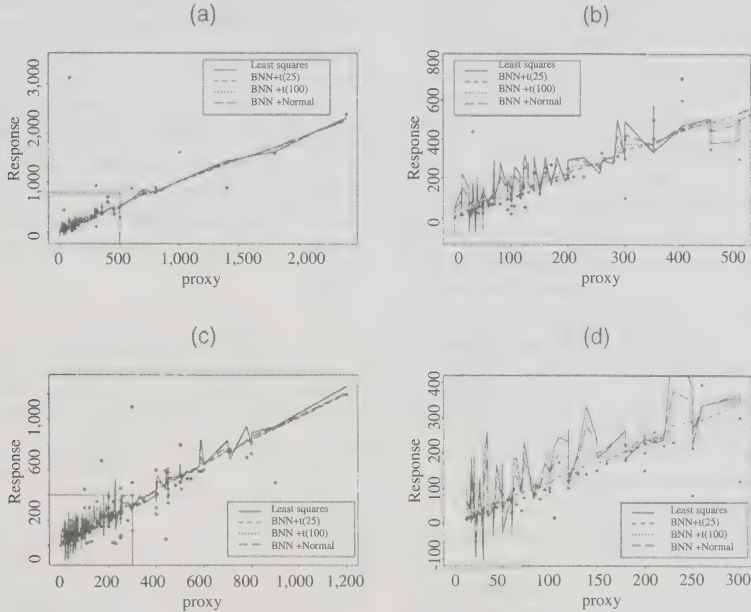


Figure 4. Fitted and predicted response curves by various models. The curves are plotted against the proxy variable, and the true response values are shown by points. (a) The fitted response curves for the sampled elements. (b) The amplification of the square region of (a). (c) The predicted response curves for the unsampled elements. (d) The amplification of the square region of (c), and for clearness only every fourth elements are plotted in the order of sorted proxy values.

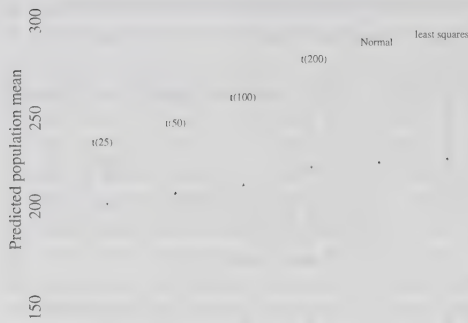


Figure 5. Estimated population mean and the associated 95% confidence interval by various models. The dotted line shows the true population mean which is 194.34.

3.4 Numerical Results on More Sample Replicates

BNN models were applied to analyze the 1,000 sample replicates. For each sample replicate of the first explanatory set, we set $\nu = 100$, $\lambda = 5$ and $M = 8$ which produces 62 connections for the full BNN model. RJEMC was run as described in section 3.3. In each run 600 BNN samples were obtained for the inference. The computational results were summarized in Table 1. It shows that BNN models have made a significantly improvement over the linear regression based models in population mean estimation for the first explanatory set. Although the BNN estimate is slightly biased (The relative bias is about 2.5% in terms of absolute values and is still acceptable.), it has the smallest MSE value among all estimates in Table 1 and the highest nominal coverage probability among the estimates with smaller MSE values (the boldfaced rows). As discussed in the last subsection, we expect \bar{y}_{BNN} to behave differently for samples containing and not containing the outlying element 53. When averaged over only those samples that contain element 53, \bar{y}_{BNN} with $\nu = 50$ performs very well with bias 1.51 and 99.6% coverage. The result is obviously not as good as for those samples not containing element 53 due to the inevitable underestimation of the finite population mean. Frankly, there is not much one can do if there are outliers in the population but none in the sample. No statistical method based on sample information alone will be able to predict the occurrence of outliers in the non-sample. We believe that \bar{y}_{BNN} will perform very well for populations without outliers due to the universal approximation property of neural networks and the technique of Bayesian model averaging.

Let \bar{x}_{11} denote the average of proxy values of the elements in one sample replicate. To see how the performance of the BNN models varied with \bar{x}_{11} , we ordered the 1,000 sample replicates according to their values of \bar{x}_{11} and

divided them into 20 groups of 50 replicates, the first group containing the 50 replicates whose \bar{x}_{11} are smallest, and so forth. For each group, we calculated MEAN, MSE and AV MSE. Figure 6 shows these conditional values. From Figure 6(a) it is easy to see that BNN models possess one good property, namely, the population mean estimate is not sensitive to the value of \bar{x}_{11} . From Figure 6(b) it is easy to see that AV MSE provides an essentially unbiased estimate for MSE regardless of averaged proxy values.

To assess the influence of ν , M and λ on BNN model size and prediction ability for the first explanatory set, we conducted three groups of experiments. In the first group of experiments, we fixed $M = 8$ and $\lambda = 5$, and varied the value of ν , $\nu = 50, 100$ and 150. In the second group of experiments, we fixed $\nu = 100$ and $\lambda = 5$, and varied the value of M , $M = 6, 8$ and 10. In the third group of experiments, we fixed $\nu = 100$ and $M = 8$, and varied the value of λ , $\lambda = 4, 5$ and 6. For each setting, RJEMC was run as described in section 3.3 for the 1,000 sample replicates. The computational results were summarized in Table 2. It shows that the averaged model size produced by each setting is about the same, although it increases slowly as M and λ increase. The results of the first group of experiments show clearly that for BNN models there is a trade-off between BIAS and MSE or AV MSE by choosing the value of ν . The results of the second and third group of experiments show that BIAS, MSE, AV MSE and the coverage probability are rather stable to the variation of M and λ , although the latter three statistics have a slow tendency to increase as M and λ increase. The increasing trend of these statistics is due to the fact that the neural networks tend to be overfitted as M and λ increase.

Table 1

Bias, mean squared error, average of mean squared error estimates and empirical coverage of various estimation strategies for the population mean using x_1, \dots, x_d and x_{11} as auxiliary variables. Figures other than BNN are reproduced from Silva and Skinner (1997).

Estimation strategy	BIAS	MSE	AVMSE	Coverage ^a (%)
SM) Sample mean (\bar{y}, V_y)	0.25	620.09	619.05	91.8
CN) Cond. num. red. (\bar{y}, V_y)	0.34	507.33	483.63	89.8
RI) Ridge	2.12	304.95	257.07	82.5
Fs) Forward (\bar{y}_r, V_y)	0.40	233.78	239.62	82.7
Fd) Forward (\bar{y}_r, V_d)	-1.25	188.08	196.88	82.0
Fg) Forward (\bar{y}_r, V_g)	-1.28	188.38	192.73	81.1
Bs) Best (\bar{y}_r, V_s)	0.44	236.90	239.49	82.7
Bd) Best (\bar{y}_r, V_d)	-1.22	190.52	196.84	82.0
Bg) Best (\bar{y}_r, V_g)	-1.24	190.83	192.71	81.1
FI) Fixed (\bar{y}_r, V_s)	0.29	227.90	241.24	83.3
SI) Saturated (\bar{y}_r, V_s)	0.30	233.58	242.32	82.5
FR) Proc REG (\bar{y}_r, V_s)	0.38	235.86	240.26	82.5
BNN) $t(100)$	-4.91	138.11	127.14	84.8

^a Nominal 95% coverage.

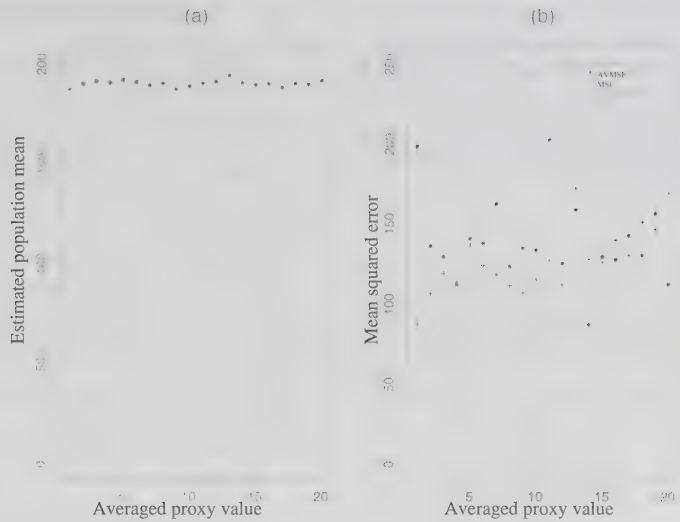


Figure 6. MEAN (panel (a)), MSE and AVMSE (Panel (b)) conditional on the averaged proxy values. The 1,000 sample replicates are ordered on \bar{x}_{11} and divided into 20 groups of 50 samples.

Table 2

Assessment of the influence of ν , M and λ on BNN model size and prediction ability for the first explanatory set. For convenience of comparison, the results of the setting $\nu = 100$, $M = 8$ and $\lambda = 5$ were repeated in panels B and C.

Experiment	ν	M	λ	Size ^a	BIAS	MSE	AVMSE	Coverage ^b (%)
A	50	8	5	10.53	-6.78	131.78	90.08	82.0
	100	8	5	10.70	-4.91	138.11	127.14	84.8
	150	8	5	10.79	-3.81	156.55	160.28	85.5
B	100	6	5	9.52	-4.90	136.72	122.58	84.1
	100	8	5	10.70	-4.91	138.11	127.14	84.8
	100	10	5	11.83	-5.14	140.13	132.20	86.4
C	100	8	4	9.42	-4.94	138.04	125.99	85.2
	100	8	5	10.70	-4.91	138.11	127.14	84.8
	100	8	6	11.83	-4.92	139.62	128.64	85.7

^a Size = $\sum_{k=1}^{1,000} \sum_{i=1}^M m(\Lambda_i) / M / 1,000$, where $m(\Lambda_i)$ is the number of connections of the neural network Λ_i .

^b Nominal 95% coverage.

The above experiments also address the issue of model misspecification. Note the BNN model proposed in this paper is specified by the three parameters, ν , M and λ . Table 2 shows that the BNN model can still perform well even when the parameter setting has some departures from the optimal setting. In practice, the setting of ν , M and λ can be determined by a cross-validation experiment. This will be demonstrated in the second simulation study.

Finally, we consider the weaker set of auxiliary variables x_1, \dots, x_{10} . For each sample replicate, we set $\nu = 100$, $\lambda = 5$ and $M = 8$ which produces 107 connections for the full BNN model. RJEMC was run as in section 3.3. The

computational results were summarized in Table 3. It shows clearly that BNN models continue to provide a significant improvement over the linear regression based models in population mean estimation when the strongest predictor x_{11} is excluded. The BNN estimate has the smallest MSE value among all estimates in Table 3, and has the smallest bias and the highest nominal coverage probability among the estimates with smaller MSE values (the boldfaced rows).

To assess the influence of ν , M and λ on BNN model sizes and prediction abilities for the second explanatory set, we conducted the same three groups of experiments as for

the first explanatory set. The computational results were summarized in Table 4. Panel A shows again the trade-off between BIAS and MSE or AVMSE made for BNN models by the value of v . Panels B and C show that BIAS, MSE, AVMSE and the coverage probability have an even more stable performance across different choices of M and λ than that of the first explanatory set.

Table 3

Bias, mean squared error, average of mean squared error estimates and empirical coverage of various estimation strategies for the population mean using x_1, \dots, x_{10} as auxiliary variables. Figures other than BNN are reproduced from Silva and Skinner (1997).

Estimation strategy	BIAS	MSE	AVMSE	Coverage ^a (%)
SM) Sample mean (\bar{y}, V_s)	0.25	620.09	619.05	91.8
CN) Cond. num. red. (\bar{y}, V_s)	3.49	562.91	450.36	87.3
RI) Ridge	1.05	480.18	472.82	89.4
Fs) Forward (\bar{y}_r, V_s)	0.06	468.46	397.99	86.7
Fd) Forward (\bar{y}_r, V_d)	-8.12	434.27	338.90	81.7
Fg) Forward (\bar{y}_r, V_g)	-7.90	433.71	328.46	81.6
Bs) Best (\bar{y}_r, V_s)	-0.00	466.16	397.59	86.6
Bd) Best (\bar{y}_r, V_d)	-7.90	434.54	336.88	81.5
Bg) Best (\bar{y}_r, V_g)	-7.60	433.26	326.05	81.6
FI) Fixed (\bar{y}_r, V_s)	0.45	490.49	461.86	89.0
SS) Saturated (\bar{y}_r, V_s)	-0.20	462.71	413.17	86.9
FR) Proc REG (\bar{y}_r, V_s)	-0.07	466.13	399.34	86.4
BNN) $t(100)$	-5.78	395.25	323.12	86.5

^a Nominal 95% coverage.

4. SECOND SIMULATION STUDY

In the first simulation study, we show that the BNN model works well for the data sets with outliers. In this simulation study, we show that the BNN model works even better for the data sets without outliers. In this study, we also demonstrate how a cross-validation procedure can be applied to determine a setting for the parameters v , M and λ of the BNN model.

The simulation population comprises the records of the serious crimes of 141 large standard Metropolitan Statistical Areas (SMSAs) in the United States. A SMSA includes a city (or cities) of specified population size. The data generally pertains to the years 1976 and 1977, and is available in Neter, Kutner, Nachtsheim and Wasserman (1996). We consider the total number of serious crimes in 1977 as the survey variable (y) and the following 9 variables as potential auxiliary variables.

x_1	Land area (in square miles);
x_2	Estimated 1977 total population (in thousands);
x_3	Percent of 1976 SMSA population in central city or cities;
x_4	Percent of 1976 SMSA population 65 years old or older;
x_5	Number of professionally active nonfederal physicians as of December 31, 1977;
x_6	Total number of beds, cribs, and bassinets during 1977;
x_7	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school, according to the 1970 Census of the Population;
x_8	Total number of persons in civilian labor force (persons 16 years old or older classified as employed or unemployed) in 1977 (in thousands);
x_9	Total current income received in 1976 by residents of the SMSA from all sources (in millions of dollars).

Table 4

Assessment of the influence of v , M and λ on BNN model size and prediction ability for the second explanatory set. For convenience of comparison, the results of the setting $v = 100$, $M = 8$ and $\lambda = 5$ were repeated in panels B and C of the table.

Experiment	v	M	λ	Size ^a	BIAS	MSE	AVMSE	Coverage ^b (%)
A	50	8	5	14.87	-9.30	394.11	270.09	82.5
	100	8	5	15.06	-5.78	395.25	323.12	86.5
	150	8	5	15.17	-4.38	412.56	346.75	87.1
B	100	6	5	13.90	-5.77	394.79	319.13	86.0
	100	8	5	15.06	-5.78	395.25	323.12	86.5
	100	10	5	16.05	-5.91	396.27	327.86	87.1
C	100	8	4	13.23	-5.62	397.65	323.68	86.4
	100	8	5	15.06	-5.78	395.25	323.12	86.5
	100	8	6	16.76	-5.78	396.45	321.98	86.6

^a Size = $\sum_{i=1}^{1,000} m(\Lambda_i) / M / 1,000$, where $m(\Lambda_i)$ is the number of connections of the neural network Λ_i .

^b Nominal 95% coverage.

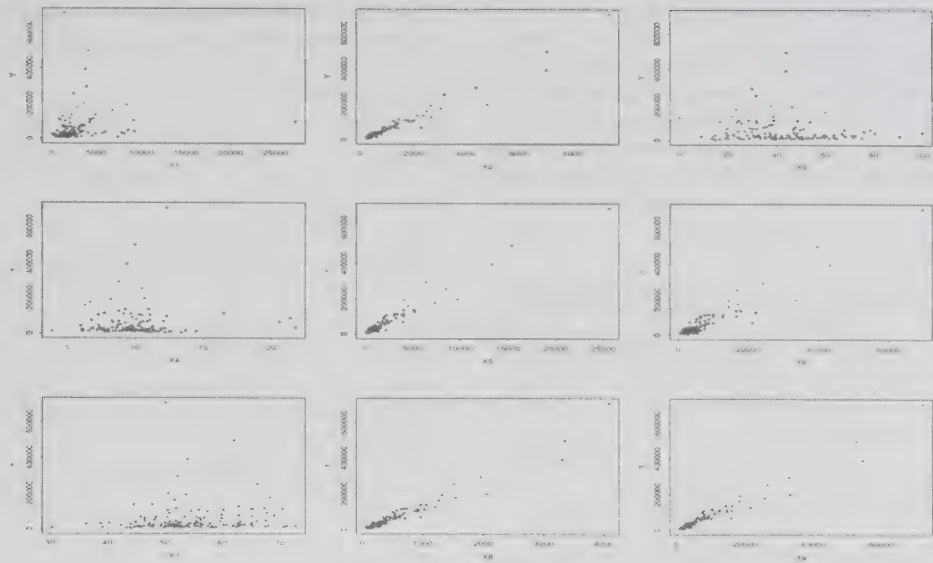


Figure 7: Scatter plots of the response variable y versus the auxiliary variables for the second simulation study.

Table 5

Cross-validation experiments for the SMSA example. For convenience of comparison, the results of the setting $\nu = 100$, $M = 3$ and $\lambda = 5$ were repeated in panels B and C.

Experiment	ν	M	λ	Size	BIAS ($\times 10^3$)	MSE ($\times 10^6$)	AVMSE ($\times 10^6$)	Coverage ^a (%)
A	50	3	5	10.68	-0.472	4.78	4.19	91
	100	3	5	10.74	-0.527	5.04	4.24	92
	∞	3	5	10.74	-0.543	4.76	4.21	92
B	100	1	5	7.29	-0.466	4.63	3.66	89
	100	2	5	9.42	-0.500	4.61	3.91	90
	100	3	5	10.74	-0.527	5.04	4.24	92
	100	4	5	11.66	-0.480	4.74	4.47	91
C	100	3	4	9.56	-0.434	4.68	4.12	92
	100	3	5	10.74	-0.527	5.04	4.24	92
	100	3	6	11.82	-0.455	4.66	4.28	93

^a Nominal 95% coverage.

Figure 7, the scatter plot of y versus the 9 auxiliary variables, suggests that a linear regression model may not be appropriate for the data set. There is a strong nonlinear relationship between y and x_1, x_3, x_4 and x_7 . Also, the explanatory variables x_2, x_5, x_6, x_8 and x_9 are highly correlated. First, we demonstrate how a cross-validation procedure can be applied to determine the setting for the parameters ν , M and λ of the BNN model. We treated the first 70 records as a small finite population, generated 100 sample replicates of size 50 from these 70 records by the method of simple random sampling without replacement, and then conducted the following experiments. In the first group of experiments, we fixed $M = 3$ and $\lambda = 5$, and varied the value of ν , $\nu = 50, 100$ and ∞ , where $\nu = \infty$ is just an

indicator which indicates the normality assumption for the disturbance. Note $M = 3$ results in a full model of 43 connections, which has been large enough for the data set. In the second group of experiments, we fixed $\nu = 100$ and $\lambda = 5$, and varied the value of M , $M = 1, 2, 3, 4$. In the third group of experiments, we fixed $\nu = 100$ and $M = 3$, and varied the value of λ , $\lambda = 4, 5, 6$. For each setting, RJEMC was run as in the first simulation study. The computational results were summarized in Table 5. It shows that the performance of the BNN model is rather stable to the variation of the settings. It also suggests that the setting $\nu = 100$, $M = 3$ and $\lambda = 4$ probably be a good setting for this simulation population by a synthetical considerations on all values of BIAS, MSE, AVMSE and coverage probability.

In the further analysis, we generated 500 sample replicates of size 70 from all the 141 records by the method of simple random sampling without replacement. For each replicate, RJEMC was run as in the first simulation study. The computational results were summarized in Table 6. It shows that the BNN model also works well for this population. We also tried the other settings given in Table 5 for the 500 sample replicates. The computational results are all similar.

Table 6

Computational results for the second simulation study with
 $v = 100, M = 3$ and $\lambda = 4$

Size	BIAS ($\times 10^3$)	MSE ($\times 10^6$)	AVMSE ($\times 10^6$)	Coverage ^a (%)
9.20	-0.512	3.36	3.25	92.6

^a Nominal 95% coverage.

5. DISCUSSION

In this article, we studied the use of Bayesian neural networks in finite population estimation. The numerical results show that it has made a significant improvement over the linear regression based methods. The improvement is not from Bayesian model averaging, but mainly from BNN models. We also applied the linear regression based Bayesian model averaging method (Liang, Truong and Wong 2001) to the same problem, and the improvement over Silva and Skinner (1997) is only marginal. Although our implementation for BNN models is not specific to finite populations, we do not think this is a shortcoming of our method. The generality of our method suggests its wide applications, for example, in nonlinear regression and nonlinear time series (the program is available by an request from the first author). Of course, a further research on how to use the known auxiliary variable information for a finite population in BNN training is also of interest.

APPENDIX

Before proving Theorem 2.1, we give one formula which will be used in the proof.

Formula 5.1 (Laplace's method)

$$\int b(\theta) \exp\{-nh(\theta)\} d\theta = (2\pi/n)^{p/2} \left| \sum \right|^{1/2} \exp\{-nh(\hat{\theta})\} b(\hat{\theta}) \{1 + O(n^{-1})\}, \quad (24)$$

as $n \rightarrow \infty$, where $b(\cdot)$ is a general function which does not depend on n , $h(\theta)$ is a constant-order function of n as $n \rightarrow \infty$, p is the dimension of θ , $\hat{\theta}$ is the maximizer of $-h(\theta)$ and $\sum = (D^2 h(\hat{\theta}))^{-1}$ is the inverse of the negative Hessian matrix evaluated at $\hat{\theta}$.

For the general formulation of Laplace's method, see Kass and Vaidyanathan (1992).

Proof of Theorem 2.1

Proof: **Part (a)** By definition of expectation, $E_{\pi} |g(\mathbf{x}_0, \theta_{\Lambda})|^{2+\delta}$ can be written as

$$E_{\pi} |g(\mathbf{x}_0, \theta_{\Lambda})|^{2+\delta} = \sum_{k=0}^K P(\Lambda_k | D) \int |g(\mathbf{x}_0, \theta_{\Lambda})|^{2+\delta} \pi(\theta_k | \Lambda_k, D) d\theta_k.$$

Following from the normality of the posterior distributions $\pi(\theta_k | \Lambda_k, D)$ (Walker 1969) and the fact that the activation function $\psi(\cdot)$ in (3) is bounded, we know (9) holds. Walker (1969) showed that the posterior distribution is Gaussian in the limit of infinite training data.

Part (b). For a given observation \mathbf{x}_0 , $E_{\pi} \hat{g}(\mathbf{x}_0, \theta_{\Lambda})$ can be written as

$$E_{\pi} = \hat{g}(\mathbf{x}_0, \theta_{\Lambda}) = \frac{\sum_{\Lambda \in \Omega} P(\Lambda) \int \hat{g}(\mathbf{x}_0, \theta_{\Lambda}) \exp\{-nh(\theta_{\Lambda})\} \tilde{\pi}(\theta_{\Lambda} | \Lambda) d\theta_{\Lambda}}{\sum_{\Lambda \in \Omega} P(\Lambda) \int \exp\{-nh(\theta_{\Lambda})\} \tilde{\pi}(\theta_{\Lambda} | \Lambda) d\theta_{\Lambda}} \quad (25)$$

where

$$\begin{aligned} \log \tilde{\pi}(\theta_{\Lambda} | \Lambda) = & -\log \sigma^2 - \frac{1}{2} \sum_{i=0}^p I_{\alpha_i} \left(\log \sigma_{\alpha}^2 + \frac{\alpha_i^2}{\sigma_{\alpha}^2} \right) \\ & - \frac{1}{2} \sum_{j=1}^M I_{\beta_j} \delta \left(\sum_{i=0}^p I_{\gamma_{ji}} \right) \left(\log \sigma_{\beta}^2 + \frac{\beta_j^2}{\sigma_{\beta}^2} \right) \\ & - \frac{1}{2} \sum_{j=1}^M \sum_{i=0}^p I_{\beta_j} I_{\gamma_{ji}} \left(\log \sigma_{\gamma}^2 + \frac{\gamma_{ji}^2}{\sigma_{\gamma}^2} \right) \\ & - \frac{m}{2} \log(2\pi) + m \log \lambda - \log(m!), \end{aligned} \quad (26)$$

and

$$\begin{aligned} h(\theta_{\Lambda}) = & \frac{1}{n} \left[\frac{n}{2} \log \sigma^2 + \frac{v+1}{2} \sum_{i=1}^n \log \left(1 + \frac{(y_i - \hat{f}(\mathbf{x}_i))^2}{v \sigma^2} \right) \right] \\ \approx & \frac{1}{n} \left[\frac{n}{2} \log \sigma^2 + \frac{v+1}{2} \sum_{i=1}^n \frac{(y_i - \hat{f}(\mathbf{x}_i))^2}{v \sigma^2} \right] \\ \approx & \frac{1}{2} \log \sigma^2 + \frac{v+1}{2v \sigma^2} E(y_i - \hat{g}(\mathbf{x}_i, \theta_{\Lambda}))^2 \\ = & \frac{1}{2} \log \sigma^2 + \frac{v+1}{2v \sigma^2} [E(y_i - g(\mathbf{x}_i))^2 + (g(\mathbf{x}_i) - \hat{g}(\mathbf{x}_i, \theta_{\Lambda}))^2], \end{aligned} \quad (27)$$

where the first approximation follows from the Taylor expansion, $\log(1+z) \approx z$, when z lies in a neighbourhood of zero; and the second approximation follows from the weak law of large numbers by assuming that n is large. Note v is often set to a large number, say, a number greater than 30. In the first example of this paper, we set $v = 100$. The equation (27) implies that the minimum of $h(\theta_\Lambda)$ is attained when $g(\mathbf{x}_i) = \hat{g}(\mathbf{x}_i, \theta_\Lambda)$ holds, that is, $\hat{g}(\mathbf{x}_i, \theta_\Lambda) = g(\mathbf{x}_i)$, where $\hat{\theta}_\Lambda = \arg \min_{\theta_\Lambda} h(\theta_\Lambda)$.

By applying Laplace's method to the numerator of (25) with $b(\cdot) = \hat{g}(\mathbf{x}_0, \theta_\Lambda) \tilde{\pi}(\theta_\Lambda | D)$, we have

$$\begin{aligned} & \sum_{\Lambda \in \Omega} P(\Lambda) \int \hat{g}(\mathbf{x}_0, \theta_\Lambda) \exp\{-\tau H(\theta_\Lambda)\} \tilde{\pi}(\theta_\Lambda | \Lambda) d\theta_\Lambda \\ & \approx \sum_{\Lambda \in \Omega} P(\Lambda) (2\pi/n)^{m/2} |\Sigma_\Lambda|^{1/2} \\ & \quad \exp\{-nh(\hat{\theta}_\Lambda)\} \hat{g}(\mathbf{x}_0, \hat{\theta}_\Lambda) \tilde{\pi}(\hat{\theta}_\Lambda | D) \\ & \approx g(\mathbf{x}_0) \sum_{\Lambda \in \Omega} P(\Lambda) (2\pi/n)^{m/2} |\Sigma_\Lambda|^{1/2} \\ & \quad \exp\{-nh(\hat{\theta}_\Lambda)\} \tilde{\pi}(\hat{\theta}_\Lambda | D), \end{aligned} \quad (28)$$

where the first approximation follows from the Laplace formula (24), and the second approximation follows from the equality $\hat{g}(\mathbf{x}_i, \hat{\theta}_\Lambda) = g(\mathbf{x}_i)$. Here we assume that the number of hidden units of each Λ is sufficiently large such that $g(\cdot)$ can be approximated arbitrarily well by the network with properly adjusted weights. Otherwise, that term will take a small value and is negligible in the last approximation of (28).

Similarly, by applying the Laplace's method to the denominator of (25) with $b(\cdot) = \tilde{\pi}(\theta_\Lambda | D)$, we have

$$\begin{aligned} & \sum_{\Lambda \in \Omega} P(\Lambda) \int \exp\{-nh(\theta_\Lambda)\} \tilde{\pi}(\theta_\Lambda | \Lambda) d\theta_\Lambda \\ & \approx \sum_{\Lambda \in \Omega} P(\Lambda) (2\pi/n)^{m/2} |\Sigma_\Lambda|^{1/2} \exp\{-nh(\hat{\theta}_\Lambda)\} \tilde{\pi}(\hat{\theta}_\Lambda | D). \end{aligned} \quad (29)$$

Following from (28), (29), and the approximation accuracy ($O(n^{-1})$) of Laplace's method, we have

$$E_\pi \hat{g}(\mathbf{x}_0, \theta_{\Lambda_i}) \rightarrow g(\mathbf{x}_0), \quad (30)$$

as $n \rightarrow \infty$. Following from (7), (9) and (30), we have

$$\frac{1}{M} \sum_{i=1}^M \hat{g}(\mathbf{x}_0, \theta_{\Lambda_i}) \rightarrow g(\mathbf{x}_0), \quad a.s.,$$

as $n \rightarrow \infty$ and $M \rightarrow \infty$.

Part (c). It follows from (8), (9), (30) and Slutsky's Theorem (Casella and Berger 2002). The proof is completed.

ACKNOWLEDGEMENTS

The authors would like to thank Chris Skinner for providing the test census data set, and thank the anonymous referees, the associate editor and editor Dr. M.P. Singh for their constructive comments which have led to a significant improvement of this paper.

REFERENCES

- BANKIER, M.D. (1990). Two step generalized least squares estimation. Ottawa: Statistics Canada, Social Survey Methods Division, Internal reports.
- BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BILLINGSLEY, P. (1986). *Probability and Measure* (Second Edition). New York: John Wiley & Sons, Inc.
- BREIDT, F.J., and OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.
- BUNTINE, W.L., and WEIGEND, A.S. (1991). Bayesian back-propagation. *Complex Systems*, 5, 603-643.
- CASELLA, G., and BERGER, R.L. (2002). *Statistical Inference* (Second Edition). United States: Thompson Learning.
- CHAMBERS, R.L., DORFMAN, A.H. and WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.
- COCHRAN, W.G. (1977). *Sampling techniques* (3rd Ed.). New York: John Wiley & Sons, Inc.
- CYBENKO, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303-314.
- DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.
- DEVILLE, J.-C., and SÄRNDALL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DORFMAN, A.H. (1992). Non-parametric regression for estimating totals in finite populations. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA. 622-625.
- DUNSTAN, R., and CHAMBERS, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.
- FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society B*, 60, 3-21.
- FUNAHASHI, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2, 183-192.
- GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-472.

- GREEN, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711-732.
- HOETING, J.A., MADIGAN, D., RAFTERY, A.E. and VOLINSKY, C. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, 14, 382-417.
- HOLMES, C.C., and MALLICK, B.K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation*, 10, 1217-1233.
- HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359-366.
- KASS, R.E., and VAIDYANATHAN, S. (1992). Approximate Bayesian factor and orthogonal parameters, with applications to testing equality of two binomial proportions. *Journal of the Royal Statistical Society B*, 54, 129-144.
- KUK, A.Y.C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80, 385-392.
- KUK, A.Y.C., and WELSH, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society B*, 63, 277-292.
- LIANG, F., TRUONG, Y.K. and WONG, W.H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statistica Sinica*, 11, 1005-1029.
- LIANG, F., and WONG, W.H. (2001). Real parameter evolutionary Monte Carlo with applications in Bayesian mixture models. *Journal of the American Statistical Association*, 96, 653-666.
- MACKAY, D.J.C. (1992). A practical Bayesian framework for backprop networks. *Neural Computation*, 4, 448-472.
- MADIGAN, D., and RAFTERY, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535-1546.
- MARRS, A.D. (1998). An application of reversible-jump MCMC to multivariate spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 10*. San Mateo, CA: Morgan Kaufmann. 577-583.
- MÜLLER, P., and INSUA, D.R. (1998). Issues in Bayesian analysis of neural network models. *Neural Computation*, 10, 749-770.
- NEAL, R.M. (1996). *Bayesian Learning For Neural Networks*. New York: Springer-Verlag.
- NETER, J., KUTNER, M.H., NACHTSHEIM, C.J. and WASSERMAN, W. (1996). *Applied Linear Statistical Models* (Fourth Edition). Chicago: Irwin.
- ROBERTS, C.P., and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer.
- ROBERTS, G.O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (Eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter). London: Chapman & Hall/CRC. 45-57.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- SILVA, P.L.D. (1996). Some asymptotic results on the mean squared error of the regression estimator under simple random sampling without replacement. Southampton: University of Southampton, Center for Survey Data Analysis Technical Report 6-2.
- SILVA, P.L.D., and SKINNER, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.
- THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal American Statistical Association*, 94, 635-644.
- TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701-1786.
- VALLIANT, R., DORFMAN, A.H. and ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- WALKER, A.M. (1969). On the asymptotic behaviour of posterior distributions. *Journal Royal Statistics Society, B*, 31, 80-88.
- WEIGEND, A.S., HUBERMAN, B.A. and RUMELHART, D.E. (1990). Predicting the future: A connectionist approach. *Int. J. Neural Syst.* 1, 193-209.
- WU, C., and SITTE, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal American Statistical Association*, 96, 185-193.

Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation

JEROME P. REITER¹

ABSTRACT

Several statistical agencies use, or are considering the use of, multiple imputation to limit the risk of disclosing respondents' identities or sensitive attributes in public use data files. For example, agencies can release partially synthetic datasets, comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. This article presents an approach for generating multiply-imputed, partially synthetic datasets that simultaneously handles disclosure limitation and missing data. The basic idea is to fill in the missing data first to generate m completed datasets, then replace sensitive or identifying values in each completed dataset with r imputed values. This article also develops methods for obtaining valid inferences from such multiply-imputed datasets. New rules for combining the multiple point and variance estimates are needed because the double duty of multiple imputation introduces two sources of variability into point estimates, which existing methods for obtaining inferences from multiply-imputed datasets do not measure accurately. A reference t -distribution appropriate for inferences when m and r are moderate is derived using moment matching and Taylor series approximations.

KEY WORDS: Confidentiality; Missing data; Public use data; Survey; Synthetic data.

1. INTRODUCTION

Many statistical agencies disseminate microdata, *i.e.*, data on individual units, in public use files. These agencies strive to release files that are (i) safe from attacks by ill-intentioned data users seeking to learn respondents' identities or attributes, (ii) informative for a wide range of statistical analyses, and (iii) easy for users to analyze with standard statistical methods. Doing this well is a difficult task. The proliferation of publicly available databases, and improvements in record linkage technologies, have made disclosures a serious threat, to the point where most statistical agencies alter microdata before release. For example, agencies globally recode variables, such as releasing ages in five year intervals or top-coding incomes above \$100,000 as "\$100,000 or more" (Willenborg and de Waal 2001); they swap data values for randomly selected units (Dalenius and Reiss 1982); or, they add random noise to continuous data values (Fuller 1993). Inevitably, these strategies reduce the utility of the released data, making some analyses impossible and distorting the results of others. They also complicate analyses for users. To analyze properly perturbed data, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach to disseminating public use data was suggested by Rubin (1993): release multiply-imputed,

synthetic datasets. Specifically, he proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these datasets to the public. These are called *fully synthetic* data sets. Releasing fully synthetic data can protect confidentiality, since identification of units and their sensitive data is nearly impossible when the values in the released data are not actual, collected values. Furthermore, with appropriate synthetic data generation and the inferential methods developed by Raghunathan, Reiter and Rubin (2003) and Reiter (2004b), it can allow data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of fully synthetic data are described by Rubin (1993), Little (1993), Fienberg, Makov and Steele (1998), Raghunathan *et al.* (2003), and Reiter (2002, 2004a).

No statistical agencies have released fully synthetic datasets as of this writing, but some have adopted a variant of the multiple imputation approach suggested by Little (1993): release datasets comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic* datasets. For example, the U.S. Federal Reserve Board protects data in the U.S. Survey of Consumer Finances by replacing monetary values at high

¹ Jerome P. Reiter, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu.

disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell 1997). The U.S. Bureau of the Census and Abowd and Woodcock (2001) protect data in longitudinal, linked data sets by replacing all values of some sensitive variables with multiple imputations and leaving other variables at their actual values. Liu and Little (2002) present a general algorithm, named SMIke, for simulating multiple values of key identifiers for selected units.

All these partially synthetic approaches are appealing because they promise to maintain the primary benefits of fully synthetic data—protecting confidentiality while allowing users to make inferences without learning complicated statistical methods or software—with decreased sensitivity to the specification of imputation models (Reiter 2003). Valid inferences from partially synthetic datasets can be obtained using the methods developed by Reiter (2003, 2004b), whose rules for combining point and variance estimates again differ from those of Rubin (1987) and also from those of Raghunathan *et al.* (2003).

The existing theory and methods for partially synthetic data do not deal explicitly with an important practical complication: in most large surveys, there are units that fail to respond to some or all items of the survey. This article presents a multiple imputation approach that handles simultaneously missing data and disclosure limitation. The approach involves two steps. First, the agency uses multiple imputation to fill in the missing data, generating m multiply-imputed datasets. Second, the agency replaces the values at risk of disclosure in each imputed dataset with r multiple imputations, ultimately releasing mr multiply-imputed datasets. This double-duty of multiple imputation requires new methods for obtaining valid inferences from the multiply-imputed datasets, which are derived here.

The paper is organized as follows. Section 2 reviews multiple imputation for missing and partially synthetic data. Section 3 presents the new methods for generating partially synthetic data and obtaining valid inferences when some survey data are missing. Section 4 shows a derivation of these methods from a Bayesian perspective, and it discusses conditions under which the resulting inferences should be valid from a frequentist perspective. Section 5 concludes with a discussion of the challenges to implementing this multiple imputation approach on genuine data, with an aim towards stimulating future research.

2. REVIEW OF MULTIPLE IMPUTATION INFERENCES

To describe multiple imputation, we use the notation of Rubin (1987). For a finite population of size N , let $I_j = 1$ if unit j is selected in the survey, and $I_j = 0$ otherwise, where

$j = 1, 2, \dots, N$. Let $I = (I_1, \dots, I_N)$. Let R_j be a $p \times 1$ vector of response indicators, where $R_{jk} = 1$ if the response for unit j to survey item k is recorded, and $R_{jk} = 0$ otherwise. Let $R = (R_1, \dots, R_N)$. Let Y be the $N \times p$ matrix of survey data for all units in the population. Let $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ be the $n \times p$ matrix of survey data for the n units with $I_j = 1$; Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let X be the $N \times d$ matrix of design variables for all N units in the population, e.g., stratum or cluster indicators or size measures. We assume that such design information is known approximately for all population units, for example from census records or the sampling frame(s). Finally, we write the observed data as $D = (X, Y_{\text{obs}}, I, R)$.

2.1 Multiple Imputation for Missing Data

The agency fills in values for Y_{mis} with draws from the Bayesian posterior predictive distribution of $(Y_{\text{mis}} | D)$, or approximations of that distribution such as those of Raghunathan, Lepkowski, Van Hoewyk and Solenberger (2001). These draws are repeated independently $l = 1, \dots, m$ times to obtain m completed data sets, $D^{(l)} = (D, Y_{\text{mis}}^{(l)})$. Multiple rather than single imputations are used so that analysts can estimate the variability due to imputing missing data.

In each imputed data set $D^{(l)}$, the analyst estimates the population quantity of interest, Q , using some estimator q , and estimates the variance of q with some estimator u . We assume that the analyst specifies q and u by acting as if each $D^{(l)}$ was in fact collected data from a random sample of (X, Y) based on the original sampling design I , i.e., q and u are complete-data estimators.

For $l = 1, \dots, m$, let $q^{(l)}$ and $u^{(l)}$ be respectively the values of q and u in data set $D^{(l)}$. Under assumptions described in Rubin (1987), the analyst can obtain valid inferences for scalar Q by combining the $q^{(l)}$ and $u^{(l)}$. Specifically, the following quantities are needed for inferences:

$$\bar{q}_m = \sum_{l=1}^m q^{(l)} / m \quad (1)$$

$$b_m = \sum_{l=1}^m (q^{(l)} - \bar{q}_m)^2 / (m-1) \quad (2)$$

$$\bar{u}_m = \sum_{l=1}^m u^{(l)} / m. \quad (3)$$

The analyst then can use \bar{q}_m to estimate Q and $T_m = (1 + 1/m)b_m + \bar{u}_m$ to estimate the variance of \bar{q}_m .

Inferences can be based on t -distributions with degrees of freedom $v_m = (m-1)(1 + \bar{u}_m / ((1+1/m)b_m))^2$.

2.2 Multiple Imputation for Partially Synthetic Data when $Y_{\text{inc}} = Y_{\text{obs}}$

Assuming no missing data, *i.e.*, $Y_{\text{inc}} = Y_{\text{obs}}$, the agency constructs partially synthetic datasets by replacing selected values from the observed data with imputations. Let $Z_j = 1$ if unit j is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \dots, Z_n)$. Let $Y_{\text{rep},i}$ be all the imputed (replaced) values in the i^{th} synthetic data set, and let Y_{nrep} be all unchanged (unreplaced) values of Y_{obs} . The $Y_{\text{rep},i}$ are assumed to be generated from the posterior predictive distribution of $(Y_{\text{rep},i} | D, Z)$, or a close approximation of it. The values in Y_{nrep} are the same in all synthetic data sets. Each synthetic data set, d_i , then comprises $(X, Y_{\text{rep},i}, Y_{\text{nrep}}, I, Z)$. Imputations are made independently $i = 1, \dots, r$ times to yield r different partially synthetic data sets, which are released to the public. Once again, multiple imputations enable analysts to account for variability due to imputation.

The values in Z can and frequently will depend on the values in D . For example, the agency may simulate sensitive variables or identifiers only for units in the sample with rare combinations of identifiers; or, the imputer may replace only incomes above \$100,000 with imputed values. To avoid bias, the imputations should be drawn from the posterior predictive distribution of Y for those units with $Z_j = 1$. Reiter (2003) illustrates the problems that can arise when imputations are not conditional on Z .

Inferences from partially synthetic datasets are based on quantities defined in Equations (1)–(3). As shown by Reiter (2003), under certain conditions the analyst can use \bar{q}_r to estimate Q and $T_p = b_r / r + \bar{u}_r$ to estimate the variance of \bar{q}_r . Inferences for scalar Q can be based on t -distributions with degrees of freedom $v_p = (r-1)(1 + \bar{u}_r / (b_r/r))^2$.

3. PARTIALLY SYNTHETIC DATA

WHEN $Y_{\text{inc}} \neq Y_{\text{obs}}$

When some data are missing, it seems logical to impute the missing and partially synthetic data simultaneously. However, imputing Y_{mis} and Y_{rep} from the same posterior predictive distribution can result in improper imputations. For an illustrative example, suppose univariate data from a normal distribution have some values missing completely at random (Rubin 1976). Further, suppose the agency seeks to replace all values larger than some threshold with imputations. The imputations for missing data can be based on a normal distribution fit using all of Y_{obs} . However, the imputations for replacements must be based on a posterior

distribution that conditions on values being larger than the threshold. Drawing Y_{mis} and Y_{rep} from the same distribution will result in biased inferences.

Imputing the Y_{mis} and Y_{rep} separately generates two sources of variability, in addition to the sampling variability in D , that the user must account for to obtain valid inferences. Neither T_m nor T_p correctly estimate the total variation introduced by the dual use of multiple imputation. The bias of each can be illustrated with two simple examples. Suppose only one value needs replacement, but there are hundreds of missing values to be imputed. Intuitively, the variance of the point estimator of Q should be well approximated by T_m , and T_p should underestimate the variance, as it is missing a b_m . On the other hand, suppose only one value is missing, but there are hundreds of values to be replaced. The variance should be well approximated by T_p , and T_m should overestimate the variance, as it includes an extra b_m .

To allow users to estimate the total variability correctly, agencies can employ a three-step procedure for generating imputations. First, the agency fills in Y_{mis} with draws from the posterior distribution for $(Y_{\text{mis}} | D)$, resulting in m completed datasets, $D^{(1)}, \dots, D^{(m)}$. Then, in each $D^{(l)}$, the agency selects the units whose values are to be replaced, *i.e.*, whose $Z_j^{(l)} = 1$. In many cases, the agency will impute values for the same units in all $D^{(l)}$ to avoid releasing any genuine, sensitive values for the selected units. We assume this is the case throughout and therefore drop the superscript l from Z . Third, in each $D^{(l)}$, the agency imputes values $Y_{\text{rep},i}^{(l)}$ for those units with $Z_j = 1$, using the posterior distribution for $(Y_{\text{rep}} | D^{(l)}, Z)$. This is repeated independently $i = 1, \dots, r$ times for $l = 1, \dots, m$, so that a total of $M = mr$ datasets are generated. Each dataset, $d_i^{(l)} = (X, Y_{\text{nrep}}, Y_{\text{mis}}^{(l)}, Y_{\text{rep},i}^{(l)}, I, R, Z)$, includes a label indicating the l of the $D^{(l)}$ from which it was drawn. These M datasets are released to the public. Releasing such nested, multiply-imputed datasets also has been proposed for handling missing data outside of the disclosure limitation context (Shen 2000; Rubin 2003).

Analysts can obtain valid inferences from these released datasets by combining inferences from the individual datasets. As before, let q be the analyst's estimator of Q , and let u be the analyst's estimator of the variance of q . We assume the analyst specifies q and u by acting as if each $d_i^{(l)}$ was in fact collected data from a random sample of (X, Y) based on the original sampling design I . For $l = 1, \dots, m$ and $i = 1, \dots, r$, let $q_i^{(l)}$ and $u_i^{(l)}$ be respectively the values of q and u in data set $d_i^{(l)}$. The following quantities are needed for inferences about scalar Q :

$$\bar{q}_M = \sum_{l=1}^m \sum_{i=1}^r q_i^{(l)} / (mr) = \sum_{l=1}^m \bar{q}^{(l)} / m \quad (4)$$

$$\begin{aligned}\bar{b}_M &= \sum_{l=1}^m \sum_{i=1}^r (q_i^{(l)} - \bar{q}^{(l)})^2 / m(r-1) \\ &= \sum_{l=1}^m b^{(l)} / m\end{aligned}\quad (5)$$

$$B_M = \sum_{l=1}^m (\bar{q}^{(l)} - \bar{q}_M)^2 / (m-1) \quad (6)$$

$$\bar{u}_M = \sum_{l=1}^m \sum_{i=1}^r u_i^{(l)} / (mr). \quad (7)$$

The $\bar{q}^{(l)}$ is the average of the point estimates in each group of datasets indexed by l , and the \bar{q}_M is the average of these averages across l . The $b^{(l)}$ is the variance of the point estimates for each group of datasets indexed by l , and the \bar{b}_M is average of these variances. The B_M is the variance of the $\bar{q}^{(l)}$ across synthetic datasets. The \bar{u}_M is the average of the estimated variances of q across all synthetic datasets.

Under conditions described in section 4, the analyst can use \bar{q}_M to estimate Q . An estimate of the variance of \bar{q}_M is:

$$T_M = (1 + 1/m) B_M - \bar{b}_M / r + \bar{u}_M. \quad (8)$$

When n , m , and r are large, inferences can be based on the normal distribution, $(Q - \bar{q}_M) \sim N(0, T_M)$. When m and r are moderate, inferences can be based on the t -distribution, $(Q - \bar{q}_M) \sim t_{v_M}(0, T_M)$, with degrees of freedom

$$v_M = \left(\frac{((1 + 1/m) B_M)^2}{(m-1) T_M^2} + \frac{(b_M/r)^2}{m(r-1) T_M^2} \right)^{-1}. \quad (9)$$

The behavior of T_M and v_M in special cases is instructive. When r is very large, $T_M \approx T_m$. This is because the $\bar{q}^{(l)} \approx q^{(l)}$, so that we obtain the results from analyzing the $D^{(l)}$. When the fraction of replaced values is small relative to the fraction of missing values, the \bar{b}_M is small relative to B_M , so that once again $T_M \approx T_m$. In both these cases, the v_M approximately equals v_m , which is Rubin's (1987) degrees of freedom when imputing missing data only. When the fraction of missing values is small relative to the fraction of replaced values, the $B_M \approx \bar{b}_M / r$, so that T_M is approximately equal to T_p with M released datasets.

4. JUSTIFICATION OF NEW COMBINING RULES

This section presents a Bayesian derivation of the inferences described in section 3 and describes conditions under which these inferences are valid from a frequentist perspective. These results make use of the theory developed

in Rubin (1987) and Reiter (2003). For the Bayesian derivation, we assume that the analyst and imputer use the same models.

Let $D^m = \{D^{(l)} : l = 1, \dots, m\}$ be the collection of all multiply-imputed datasets before any observed values are replaced. For each $D^{(l)}$, let $q^{(l)}$ and $u^{(l)}$ be the posterior mean and variance of Q . As in Rubin (1987, Chapter 3), let B_∞ be the variance of the $q^{(l)}$ obtained when $m = \infty$.

Let $d^M = \{d_i^{(l)} : i = 1, \dots, r ; l = 1, \dots, m\}$ be the collection of all released synthetic datasets. For each $d_i^{(l)}$, let $q_i^{(l)}$ be the posterior mean of $q^{(l)}$. For each l , let $B^{(l)}$ be the variance of the $q_i^{(l)}$ obtained when $r = \infty$. Lastly, let B be the average of the $B^{(l)}$ obtained when $m = \infty$.

Using these quantities, the posterior distribution for $(Q | d^M)$ can be decomposed as

$$\begin{aligned}f(Q | d^M) &= \int f(Q | d^M, D^m, B_\infty, B) \\ &\quad f(D^m, B_\infty | d^M, B) \\ &\quad f(B | d^M) dD^m dB_\infty dB.\end{aligned}\quad (10)$$

The integration is over the distributions of the values in D that are missing and the values in each $D^{(l)}$ that are replaced with imputations; the observed, unaltered values remain fixed. We assume standard Bayesian asymptotics hold, so that complete-data inferences for Q can be based on normal distributions.

4.1 Evaluating $f(Q | d^M, D^m, B_\infty, B)$

Given D^m , the synthetic data are irrelevant, so that $f(Q | d^M, D^m, B_\infty, B) = f(Q | d^M, B_\infty)$. This is the posterior distribution of Q for multiple imputation for missing data, conditional on B_∞ . As shown by Rubin (1987), this posterior distribution is approximately

$$(Q | D^m, B_\infty) \sim N(\bar{q}_m, (1 + 1/m) B_\infty + \bar{u}_m) \quad (11)$$

where \bar{q}_m and \bar{u}_m are defined as in (1) and (3). In multiple imputation for missing data, we integrate (11) over the posterior distribution of $(B_\infty | D^m)$. This is not done here, since we integrate over $(B_\infty | d^M)$.

4.2 Evaluating $f(D^m, B_\infty | d^M, B) f(B | d^M)$

Since the distribution for Q in (11) relies only on \bar{q}_m , \bar{u}_m , and B_∞ , it is sufficient for $f(D^m, B_\infty | d^M, B)$ to determine

$$\begin{aligned}f(\bar{q}_m, \bar{u}_m, B_\infty | d^M, B) &= \\ f(\bar{q}_m, \bar{u}_m | d^M, B_\infty, B) f(B_\infty | d^M, B).\end{aligned}$$

Following Reiter (2003), we first assume replacement imputations are made so that, for all i , the sampling distributions of each $q_i^{(l)}$ and $u_i^{(l)}$ are,

$$(q_i^{(l)} | D^{(l)}, B^{(l)}) \sim N(q^{(l)}, B^{(l)}) \quad (12)$$

$$(u_i^{(l)} | D^{(l)}, B^{(l)}) \sim (u^{(l)}, << B^{(l)}). \quad (13)$$

Here, the notation $F \sim (G, << H)$ means that the random variable F has a distribution with expectation of G and variability much less than H . In actuality, $u_i^{(l)}$ is typically centered at a value larger than $u^{(l)}$, since synthetic data incorporate uncertainty due to drawing values of the parameters. For large sample sizes n , this bias should be minimal. The assumption that $E(q_i^{(l)} | D^{(l)}, B^{(l)}) = q^{(l)}$ and the normality assumption should be reasonable when the imputations are drawn from correct posterior predictive distributions, $f(Y_{\text{rep}} | D^{(l)}, Z)$, and the usual asymptotics hold.

Assuming flat priors for all $q^{(l)}$ and $v^{(l)}$, standard Bayesian theory implies that

$$(q^{(l)} | d^M, B^{(l)}) \sim N(\bar{q}^{(l)}, B^{(l)}/r) \quad (14)$$

$$(u^{(l)} | d^M, B^{(l)}) \sim (\bar{u}^{(l)}, << B^{(l)}/r) \quad (15)$$

$$\left(\frac{(r-1)b^{(l)}}{B^{(l)}} | d^M, B^{(l)} \right) \sim \chi_{r-1}^2 \quad (16)$$

where $b^{(l)}$ is defined in (5). We next assume that $B^{(l)} = B$ for all l . This should be reasonable, since the variability in posterior variances tends to be of smaller order than the variability of posterior means. Averaging across l , we obtain

$$(\bar{q}_M | d^M, B) \sim N(\bar{q}_M, B/rm) \quad (17)$$

$$(\bar{u}_M | d^M, B) \sim (\bar{u}_M, << B/rm) \quad (18)$$

where \bar{q}_M is defined in (4) and \bar{u}_M is defined in (7). The posterior distribution of $(B_\infty | d^M, B)$ is

$$\left(\frac{(m-1)B_M}{B_\infty + B/r} | d^M, B \right) \sim \chi_{m-1}^2 \quad (19)$$

where B_M is defined in (6).

Finally, the posterior distribution of $(B | d^M)$ is

$$\left(\frac{m(r-1)\bar{b}_M}{B} | d^M \right) \sim \chi_{m(r-1)}^2 \quad (20)$$

where \bar{b}_M is defined in (5).

4.3 Evaluating $f(Q | d^M)$

We need to integrate the product of (11) and (17) with respect to the distributions in (19) and (20). This can be

done by numerical integration, but it is desirable to have simpler approximations for users.

For large m and r , we can replace the terms in the variance with their approximate expectations: the $B_\infty \approx B_M - B/r$, and the $B \approx \bar{b}_M$. Hence, for large m and r , the posterior distribution of Q is approximately:

$$\begin{aligned} & (Q | d^M) \\ & \sim N(\bar{q}_M, (1+1/m)(B_M - \bar{b}_M/r) + \bar{b}_M/mr + \bar{u}_M) \\ & = N(\bar{q}_M, (1+1/m)B_M - \bar{b}_M/r + \bar{u}_M) \\ & = N(\bar{q}_M, T_M). \end{aligned} \quad (21)$$

When m and r are moderately sized, the normal distribution may not be a good approximation. To derive an approximate reference t -distribution, we use the strategies of Rubin (1987) and Barnard and Rubin (1999). That is, we assume that for some degrees of freedom v_M to be estimated,

$$\left(\frac{v_M T_M}{\bar{u}_M + (1+1/m)B_\infty + B/mr} | d^M \right) \sim \chi_{v_M}^2 \quad (22)$$

so that we can use a t -distribution with v_M degrees of freedom for inferences about Q . We approximate v_M by matching the first two moments of (22) to those of a chi-squared distribution. The details showing that v_M is approximated by the expression in (9) are provided in the appendix.

The inferences based on (4) – (9) have valid frequentist properties under certain conditions. First, the analyst must use randomization-valid estimators, q and u . That is, when q and u are applied on D to get q_{obs} and u_{obs} , the $(q_{\text{obs}} | X, Y) \sim N(Q, U)$ and $(u_{\text{obs}} | X, Y) \sim (U, << U)$, where the relevant distribution is that of I . Second, the imputations for missing data must be proper in the sense of Rubin (1987, Chapter 4). Essentially, this requires that inferences from the imputations for missing data be randomization-valid for q_{obs} and u_{obs} , under the posited non-response mechanism. Third, the imputations for partially synthetic data must be synthetically proper in the sense of Reiter (2003). This requires that the inferences from the replacement imputations associated with each $D^{(l)}$ be randomization valid for the $q^{(l)}$ and $u^{(l)}$.

In general, it is difficult to verify that imputations for missing data are proper in complex samples (Binder and Sun 1996). They may be proper for some analyses but not for others. As a result, some confidence intervals centered on unbiased estimators may not have nominal coverage rates; see Meng (1994) for a discussion of this issue. These difficulties exist for the multiple imputation approach used here, and indeed may be compounded because of the additional imputation of synthetic data.

5. CONCLUDING REMARKS

There are many challenges to using partially synthetic data approaches for disclosure limitation. Most important, agencies must decide which values to replace with imputations. General candidates for replacement include the values of identifying characteristics for units that are at high risk of identification, such as sample uniques and duplicates, and the values of sensitive variables in the tails of the distributions. Confidentiality can be protected further by, in addition, replacing values at low disclosure risk (Liu and Little 2002). This increases the variation in the replacement imputations, and it obscures any information that can be gained just from knowing which data were replaced. As with any disclosure limitation method (Duncan, Keller-McNulty and Stokes 2001), these decisions should consider tradeoffs between disclosure risk and data utility. Guidance on selecting values for replacement is a high priority for research in this area.

There remain disclosure risks in partially synthetic data no matter which values are replaced. Users can utilize the released, unaltered values to facilitate disclosure attacks, for example via matching to external databases, or they may be able to estimate actual values of Y_{obs} from the synthetic data with reasonable accuracy. For instance, if all people in a certain demographic group have the same, or even nearly the same, value of an outcome variable, the imputation models likely will generate that value for imputations. Imputers may need to coarsen the imputations for such people. As another example, when users know that a certain record has the largest value of some Y_{obs} , that record can be identified when its value is not replaced.

On the data utility side, the main challenge is specifying imputation models, both for the missing and replaced data, that give valid results. For missing data, it is well known that implausible imputation models can produce invalid inferences, although this is less problematic when imputing relatively small fractions of missing data (Rubin 1987; Meng 1994). There is an analogous issue for partially synthetic data. When large fractions of data are replaced, for example entire variables, analyses involving the replaced values reflect primarily the distributional assumptions implicit in the imputation models. When these assumptions are implausible, the resulting analyses can be invalid. Again, this is less problematic when only small fractions of values are replaced, as might be expected in many applications of the partially synthetic approach.

Certain data characteristics can be especially challenging to handle with partially synthetic data. For example, it may be desirable to replace extreme values in skewed distributions, such as very large incomes. Information about the tails of these distributions may be limited, making it difficult to draw reasonable replacements while protecting

confidentiality. As another example, randomly drawn imputations for highly structured data may be implausible, for instance unlikely combinations of family members' ages or marital statuses. These difficulties, coupled with the general limitations of inferences based on imputations, point to an important issue for research: developing and evaluating methods for generating partially synthetic data, including semi-parametric and non-parametric approaches.

We note that building the synthetic data models is generally an easier task than building the missing data models. Agencies can compare the distributions of the synthetic data to those of the observed data being replaced. When the synthetic distributions are too dissimilar from the observed ones, the imputation models can be adjusted. There usually is no such check for the missing data models.

It is, of course, impossible for agencies to anticipate every possible use of the released data, and hence impossible to generate models that provide valid results for every analysis. A more modest and attainable goal is to enable analysts to obtain valid inferences using standard methods and software for a wide range of standard analyses, such as some linear and logistic regressions. Agencies therefore should provide information that helps analysts decide what inferences can be supported by the released data. For example, agencies can include descriptions of the imputation models as attachments to public releases of data. Users whose analyses are not supported by the data may have to apply for special access to the observed data. Agencies also need to provide documentation for how to use the nested data sets. Rules for combining point estimates from the multiple data sets are simple enough to be added to standard statistical software packages, as has been done already for Rubin's (1987) rules in SAS, Stata, and S-Plus.

As constructed, the multiple imputation approach does not calibrate to published totals. This could make some users unhappy with or distrust the released data. It is not clear how to adapt the method—or, for that matter, many other disclosure limitation techniques that alter the original data—for calibration.

Missing data and disclosure risk are major issues confronting organizations releasing data to the public. The multiple imputation approach presented here is suited to handle both simultaneously, providing users with rectangular completed datasets that can be analyzed with standard statistical methods and software. There are challenges to implementing this approach in genuine applications, but, as noted by Rubin (1993) in his initial proposal, the potential payoffs of this use of multiple imputation are high. The next item on the research agenda is to investigate how well the theory works in practice, including comparisons of this approach with other disclosure limitation methods. These comparisons should

focus on measures of disclosure risks, obtained by simulating intruder behavior, and on measures of data utility for estimands of interest to users, including properties of point and interval estimates.

APPENDIX: DERIVATION OF APPROXIMATE DEGREES OF FREEDOM

Inferences from datasets with multiple imputations for both missing data and partially synthetic replacements are made using a t -distribution. A key step is to approximate the distribution of

$$\left(\frac{v_M T_M}{\bar{u}_M + (1+1/m)B_\infty + B/mr} \mid d^M \right) \quad (23)$$

as a chi-squared distribution with v_M degrees of freedom. The v_M is determined by matching the mean and variance of the inverted χ^2 distribution to the mean and variance of (23).

Let $\alpha = (B_\infty + B/r)/B_M$, and let $\gamma = B/\bar{b}_M$. Then, $(\alpha^{-1} \mid d^M, B)$ and $(\gamma^{-1} \mid d^M)$ have mean square distributions with degrees of freedom $m-1$ and $m(r-1)$, respectively. Let $f = (1+1/m)B_M/\bar{u}_M$, and let $g = (1/r)\bar{b}_M/\bar{u}_M$. We can write (23) as

$$\frac{T_M}{\bar{u}_M + (1+1/m)B_\infty + B/mr} = \frac{\bar{u}_M(1+f-g)}{\bar{u}_M(1+\alpha f - \gamma g)}. \quad (24)$$

To match moments, we need to approximate the expectation and variance of (24).

For the expectation, we use the fact that

$$\begin{aligned} E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \mid d^M\right) \\ = E\left(E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \mid d^M, B\right) \mid d^M\right). \end{aligned} \quad (25)$$

We approximate these expectations using first order Taylor series expansion in α^{-1} and γ^{-1} around their expectations, which equal one. As a result,

$$\begin{aligned} E\left(E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \mid d^M, B\right) \mid d^M\right) \\ \approx E\left(\frac{1+f-g}{1+f-\gamma g} \mid d^M\right) \approx 1. \end{aligned} \quad (26)$$

For the variance, we use the conditional variance representation

$$\begin{aligned} E\left(\text{Var}\left(\frac{1+f-g}{1+\alpha f - \gamma g} \mid d^M, B\right) \mid d^M\right) \\ + \text{Var}\left(E\left(\frac{1+f-g}{1+\alpha f - \gamma g} \mid d^M, B\right) \mid d^M\right). \end{aligned} \quad (27)$$

For the interior variance and expectation, we use a first order Taylor series expansion in α^{-1} around its expectation. Since $\text{Var}(\alpha^{-1} \mid d^M, B) = 2/(m-1)$, the expression in (27) equals approximately

$$\begin{aligned} E\left(\frac{2(1+f-g)^2 f^2}{(m-1)(1+f-\gamma g)^4} \mid d^M\right) \\ + \text{Var}\left(\frac{1+f-g}{1+f-\gamma g} \mid d^M\right). \end{aligned} \quad (28)$$

We now use first order Taylor series expansions in γ^{-1} around its expectation to determine the components of (28). The first term in (28) is,

$$\begin{aligned} E\left(\frac{2(1+f-g)f^2}{(m-1)(1+f-\gamma g)^4} \mid d^M\right) \\ \approx \frac{2f^2}{(m-1)(1+f-g)^2}. \end{aligned} \quad (29)$$

Since $\text{Var}(\gamma^{-1} \mid d^M) = 2/(m(r-1))$, the second term in (28) is

$$\begin{aligned} \text{Var}\left(\frac{1+f-g}{1+f-\gamma g} \mid d^M\right) \\ \approx \frac{2g^2}{m(r-1)(1+f-g)^2}. \end{aligned} \quad (30)$$

Combining (29) and (30), the variance of (23) equals approximately

$$\begin{aligned} \frac{2f^2}{(m-1)(1+f-g)^2} \\ + \frac{2g^2}{m(r-1)(1+f-g)^2}. \end{aligned} \quad (31)$$

Since a mean square random variable has variance equal to 2 divided by its degrees of freedom, we conclude that

$$\begin{aligned} v_M = \\ \left(\frac{f^2}{(m-1)(1+f-g)^2} + \frac{g^2}{m(r-1)(1+f-g)^2} \right)^{-1}. \end{aligned} \quad (32)$$

ACKNOWLEDGEMENTS

This research was supported by the U.S. Bureau of the Census under a contract through Datametrics Research. The author thanks Rod Little, Trivellore Raghunathan, Don Rubin, Laura Zayatz, and the referees for inspiration, guidance, and helpful comments on this topic.

REFERENCES

- ABOWD, J.M., and WOODCOCK, S.D. (2001). Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, (Eds. P. Doyle, J. Lane, L. Zayatz and J. Theeuwes), Amsterdam: North-Holland. 215–277.
- BARNARD, J., and RUBIN, D.B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.
- BINDER, D.A., and SUN, W. (1996). Frequency valid multiple imputation for surveys with a complex design. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281–286.
- DALENIUS, T., and REISS, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73–85.
- DUNCAN, G.T., KELLER-MCNULTY, S.A. and STOKES, S.L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Technical report, U.S. National Institute of Statistical Sciences.
- FIENBERG, S.E., MAKOV, U.E. and STEELE, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14, 485–502.
- FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383–406.
- KENNICKELL, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques*, (Eds. W. Alvey and B. Jamerson), Washington, D.C.: National Academy Press, 248–267.
- LITTLE, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407–426.
- LIU, F., and LITTLE, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2133–2138.
- MENG, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558–573). *Statistical Science*, 9, 538–558.
- RAGHUNATHAN, T.E., LEPKOWSKI, J.M., VAN HOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27, 85–96.
- RAGHUNATHAN, T.E., REITER, J.P. and RUBIN, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1–16.
- REITER, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18, 531–544.
- REITER, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181–189.
- REITER, J.P. (2004a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*. Forthcoming.
- REITER, J.P. (2004b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*. Forthcoming.
- RUBIN, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581–592.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- RUBIN, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462–468.
- RUBIN, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57, 3–18.
- SHEN, Z. (2000). *Nested Multiple Imputation*. Ph. D. thesis, Harvard University, Dept. of Statistics.
- WILLENBORG, L., and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2004.

- M. Axelson, *Statistics Sweden*
 J.-F. Beaumont, *Statistics Canada*
 D.R. Bellhouse, *University of Western Ontario*
 P. Biemer, *Research Triangle Institute*
 D.A. Binder, *Statistics Canada*
 J.M. Brick, *Westat, Inc.*
 J. Breidt, *Iowa State University*
 D. Cantor, *Westat, Inc.*
 P. Cantwell, *U.S. Bureau of the Census*
 A. Chaudhuri, *Institute of Engineering & Technology Lucknow*
 B.-C. Chen, *U.S. Census Bureau*
 J. Chen, *University of Waterloo*
 K.R. Copeland, *U.S. Bureau of Labor Statistics*
 J.R. Chromy, *RTI, Inc.*
 P. Dick, *Statistics Canada*
 J. Droitcour, *United States General Accounting Office*
 J.L. Eltinge, *U.S. Bureau of Labor Statistics*
 D. Fogel, *Natural Selection, Inc.*
 O. Frank, *Stockholm University*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistics Canada*
 L. Granquist, *Statistics Sweden*
 B. Graubard, *National Cancer Institute*
 D. Haziza, *Statistics Canada*
 D. Hedeker, *University of Illinois*
 D. Hedin, *Statistics Sweden*
 M.A. Hidioglou, *Office for National Statistics*
 M. Houbiers, *Statistics Denmark*
 J. Jiang, *University of California at Davis*
 D. Judkins, *Westat, Inc.*
 G. Kalton, *Westat, Inc.*
 M. Kalzoff, *National Centre for Health Statistics*
 A. Kennickell, *Federal Research Board*
 P. Kott, *USDA/NASS*
 M. Kovačević, *Statistics Canada*
 J. Kovar, *Statistics Canada*
 M.D. Larsen, *Iowa State University*
 P. Lahiri, *JPSM, University of Maryland*
 P. Lavallée, *Statistics Canada*
 R. Lehtonen, *University of Jyväskylä*
 S. Lohr, *Arizona State University*
 D. Malec, *United States Bureau of the Census*
 H. Mantel, *Statistics Canada*
 D. Marker, *Westat, Inc.*
 S.M. Miller, *U.S. Bureau of Labour Statistics*
 J. Moore, *U.S. Bureau of the Census*
 R. Mukerjee, *Indian Institute of Management*
 G. Nathan, *Hebrew University*
 D. Norris, *Statistics Canada*
 J. Opsomer, *Iowa State University*
 D. Paton, *Statistics Canada*
 D. Pfeffermann, *Hebrew University*
 J.N.K. Rao, *Carleton University*
 T.J. Rao, *Indian Statistical Institute*
 J. Reiter, *Duke University*
 L.-P. Rivest, *Université Laval*
 L. Rizzo, *Westat, Inc.*
 R.A. Rottach, *U.S. Bureau of the Census*
 K. Rust, *Westat, Inc.*
 N. Schenker, *National Center for Health Statistics*
 F.J. Scheuren, *National Opinion Research Center*
 I. Şchiopu-Kratina, *Statistics Canada*
 N. Shlomo, *Central Bureau of Statistics*
 A.C. Singh, *RTI, Inc.*
 B. Sinha, *ISI*
 M.D. Sinclair, *Mathematica Policy research*
 C.J. Skinner, *University of Southampton*
 P. Smith, *CDC*
 E. Stasny, *Ohio State University*
 D. Steel, *University of Wollongong*
 L. Stokes, *Southern Methodist University*
 C. Swartz, *Simon Fraser University*
 A. Tersine, Jr., *United States Bureau of Labor Statistics*
 A. Thivierge, *Statistics Canada*
 M. Thompson, *University of Waterloo*
 Y. Tillé, *Université de Neuchâtel*
 C. Tucker, *United States Bureau of Labor Statistics*
 J. van der Brakel, *Central Bureau of Statistics*
 R. Valliant, *JPSM, University of Michigan*
 V. Vehovar, *University of Ljubljana*
 J. Waksberg, *Westat, Inc.*
 J. Wang, *Merck Research Labs, Merck & Co., Inc.*
 F. Wein, *Federal Statistical Office*
 K.M. Wolter, *Iowa State University*
 P. Wong, *Statistics Canada*
 C. Wu, *University of Waterloo*
 Y. You, *Statistics Canada*
 R. Yucel, *University of Massachusetts*
 W. Yung, *Statistics Canada*
 A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 2004 issues: Anne-Marie Fleury, Francine Pilon-Renaud and Roberto Guido (Dissemination Division), Philippe Laroche (Marketing Division) and François Beaudin (Official Languages and Translation Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier, and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

Erratum:

In the June 2004 issue, we published a paper by D.N. Da Silva and Jean D. Opsomer on "Properties of the Weighting Cell Estimator Under a Nonparametric Response Mechanism" (pages 45-55). We would like to apologize for having incorrectly spelled out Dr. Da Silva's name. It should have read D. Nobrega Da Silva. Please note also that the corrected version appears on Statistics Canada's Web site.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 20, No. 2, 2004

Preface.....	141
Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study James L. Esposito	143
Calendar and Question-List Survey Methods: Association Between Interviewer Behaviors and Data Quality Robert F. Belli, Eun Ha Lee, Frank P. Stafford, and Chia-Hung Chou	185
A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing Adriaan W. Hoogendoorn	219
TADEQ: A Tool for the Documentation and Analysis of Electronic Questionnaires Jelke Bethlehem and Anco Hundepool	233
Relating Respondent-Generated Intervals Questionnaire Design to Survey Accuracy and Response Rate S. James Press and Judith M. Tanur	265
Developing Bilingual Questionnaires: Experiences from New Zealand in the Development of the 2001 Māori Language Survey Lyn Potaka and Suzanne Cochrane	289
The Time-line as a Device to Enhance Recall in Standardized Research Interviews: A Split Ballot Study Wander van der Vaart	301
Using Vignettes in Cognitive Research on Establishment Surveys Rebecca L. Morrison, Kristin Stettler, and Amy E. Anderson	319
Pre-printing Effects in Official Statistics: An Experimental Study Anders Holmberg	341
Exploring Confidentiality Issues Related to Dependent Interviewing: Preliminary Findings Joanne Pascale and Thomas S. Mayer	357
How Good is Good? Comparing Numerical Ratings of Response Options for Two Versions of the Self-Assessed Health Status Question Barbara Foley Wilson, Barbara M. Altman, Karen Whitaker, and Mario Callegaro	379
Identifying and Reducing Response Burdens in Internet Business Surveys Gustav Haraldsen	393
Book and Software Reviews.....	411

All inquiries about submissions and subscriptions should be directed to jos@scb.se

CONTENTS

TABLE DES MATIÈRES

Volume 32, No. 1, March/mars 2004, 1-104

Douglas P. WIENS: Éditorial / Editorial.....	1
Richard A. LOCKHART Report from the former Editor/Rapport du rédacteur en chef sortant.....	3
John E. KOLASSA Approximate multivariate conditional inference using the adjusted profile likelihood.....	5
Changbao WU Combining information from multiple surveys through the empirical likelihood method	15
Lang WU Nonlinear mixed-effect models with nonignorably missing covariates.....	27
Brajendra C. SUTRADHAR & Patrick J. FARRELL Analyzing multivariate longitudinal binary data: a generalized estimating equations approach.....	39
Mingyao AI & Runchu ZHANG Theory of optimal blocking of nonregular factorial designs	57
Fernando A. QUINTANA & Peter MÜLLER Optimal sampling for repeated binary measurements	73
Mary C. MEYER & Michael WOODROOFE Consistent maximum likelihood estimation of a unimodal density using shape restrictions.....	85
Acknowledgement of referees' services/Remerciements aux membres des jurys	101
Forthcoming Papers/Articles à paraître	102
Volume 32 (2004): Subscription rates/Frais d'abonnement	103

Volume 32, No. 2, June/juin 2004, 105-208

Florentina BUNEA & Marten H. WEGKAMP Two-stage model selection procedures in partially linear regression	105
Ao YUAN & Bertrand CLARKE Asymptotic normality of the posterior given a statistic.....	119
Malay GHOSH, James V. ZIDEK, Tapabrata MAITI & Rick WHITE The use of the weighted likelihood in the natural exponential families with quadratic variance	139
Xinsheng LIU & Jinde WANG Testing the equality of multinomial populations ordered by increasing convexity under the alternative.....	159
Eva CANTONI A robust approach to longitudinal data analysis.....	169
Cristina BUTUCEA Deconvolution of supersmooth densities with smooth noise.....	181
Michael D. PERLMAN & Sanjay CHAUDHURI The role of reversals in order-restricted inference.....	193
Arthur COHEN & Harold B. SACKROWITZ A discussion of some inference issues in order restricted models.....	199
Forthcoming Papers/Articles à paraître	206
Volume 32 (2004): Subscription rates/Frais d'abonnement	207

Volume 32, No. 3, September/septembre 2004, 209-334

Robert GENTLEMAN Some perspectives on statistical computing	209
Ryan GILL Maximum likelihood estimation in generalized broken-line regression.....	227
Stéphane HERITIER & Elvezio RONCHETTI Robust binary regression with continuous outcomes	239
Naomi S. ALTMAN & Julio C. VILLARREAL Self-modelling regression for longitudinal data with time-invariant covariates.....	251
Hubert WONG & Bertrand CLARKE Improvement over Bayes prediction in small samples in the presence of model uncertainty	269
Tim SWARTZ, Yoel HAITOVSKY, Albert VEXLER & Tae YANG Bayesian identifiability and misclassification in multinomial data	285
Chang Xuan MAO & Bruce G. LINDSAY Estimating the number of classes in multiple populations: a geometric analysis	303
Ramon OLLER, Guadalupe GÓMEZ & M. Luz CALLE Interval censoring: model characterizations for the validity of the simplified likelihood	315
Jerald F. LAWLESS A note on interval-censored lifetime data and the constant-sum condition of Oller, Gómez & Calle (2004)	327
Forthcoming Papers/Articles à paraître	332
Volume 33 (2005): Subscription rates/Frais d'abonnement	333

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1.

Présentation
- 1.1

Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2

Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3

Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4

Les remerciements doivent paraître à la fin du texte.
- 1.5

Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2.

Résumé
- Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3.

Rédaction
- 3.1

Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2

Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
- 3.3

Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4

Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5

Distinguer clairement les caractères ambigus (comme w, ; o, O, 0, j, l).
- 3.6

Les caractères italiques sont utilisés pour faire ressortir des mots.
4.

Figures et tableaux
- 4.1

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2

Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).
5.

Bibliographie
- 5.1

Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2

La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Volume 32, No. 3, September/septembre 2004, 209-334

Robert GENTLEMAN	Some perspectives on statistical computing	209
Ryan GILL	Maximum likelihood estimation in generalized broken-line regression	227
Stéphane HERITIER & Elvezio RONCHETTI	Robust binary regression with continuous outcomes	239
Naomi S. ALTMAN & Julio C. VILLARREAL	Self-modelling regression for longitudinal data with time-invariant covariates	251
Hubert WONG & Bertrand CLARKE	Improvement over Bayes prediction in small samples in the presence of model uncertainty	269
Tim SWARTZ, Yoel HAITOVSKY, Albert VEXLER & Tae YANG	Bayesian identifiability and misclassification in multinomial data	285
Chang Xuan MAO & Bruce G. LINDSAY	Estimating the number of classes in multiple populations: a geometric analysis	303
Ramon OLLER, Guadalupe GÓMEZ & M. Luz CALLE	Interval censoring: model characterizations for the validity of the simplified likelihood	315
Jerald F. LAWLESS	A note on interval-censored lifetime data and the constant-sum condition of Oller, Gómez & Calle (2004)	327
	Forthcoming Papers/Articles à paraître	332
	Volume 33 (2005): Subscription rates/Frais d'abonnement	333

Volume 32, No. 1, March/mars 2004, 1-104

Volume 32, No. 2, June/juin 2004, 105-208

1	Douglas P. WIENS: Editorial / Editorial.....
3	Richard A. LOCKHART Report from the former Editor/Rapport du rédacteur en chef sortant.....
5	John E. KOLASSA Approximate multivariate conditional inference using the adjusted profile likelihood.....
15	Changbao WU Combining information from multiple surveys through the empirical likelihood method.....
27	Lang WU Nonlinear mixed-effect models with nonignorably missing covariates.....
39	Bragendra C. SUTRADHAR & Patrick J. FARRELL Analyzing multivariate longitudinal binary data: a generalized estimating equations approach.....
57	Mingyao AI & Runchu ZHANG Theory of optimal blocking of nonregular factorial designs.....
73	Fernando A. QUINTANA & Peter MÜLLER Optimal sampling for repeated binary measurements.....
85	Mary C. MEYER & Michael WOODROOFE Consistent maximum likelihood estimation of a unimodal density using shape restrictions.....
101	Acknowledgement of referees' services/Remerciements aux membres des jurys.....
102	Forthcoming Papers/Articles à paraître.....
103	Volume 32 (2004): Subscription rates/Frais d'abonnement.....
119	Ao YUAN & Bertrand CLARKE Asymptotic normality of the posterior given a statistic.....
139	Malay GHOSH, James V. ZIDEK, Tapabrata MAITI & Rick WHITE The use of the weighted likelihood in the natural exponential families with quadratic variance.....
159	Xinsheng LIU & Jinde WANG Testing the equality of multinomial populations ordered by increasing convexity under the alternative.....
169	Eva CANTONI A robust approach to longitudinal data analysis.....
181	Christina BUTTCEA Deconvolution of supersmooth densities with smooth noise.....
193	Michael D. PERLMAN & Sanjay CHAUDHURI The role of reversals in order-restricted inference.....
199	Arthur COHEN & Harold B. SACKROWITZ A discussion of some inference issues in order restricted models.....
206	Forthcoming Papers/Articles à paraître.....
207	Volume 32 (2004): Subscription rates/Frais d'abonnement.....

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 20, No. 2, 2004

Preface.....	141
Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study James L. Esposito	143
Calendar and Question-List Survey Methods: Association Between Interviewer Behaviors and Data Quality Robert F. Belli, Eun Ha Lee, Frank P. Stafford, and Chia-Hung Chou.....	185
A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing Adriaan W. Hoogendoorn.....	219
TADEQ: A Tool for the Documentation and Analysis of Electronic Questionnaires Jelke Bethlehem and Anco Hundepool.....	233
Relating Respondent-Generated Intervals Questionnaire Design to Survey Accuracy and Response Rate S. James Press and Judith M. Tanur.....	265
Developing Bilingual Questionnaires: Experiences from New Zealand in the Development of the 2001 Maori Language Survey Lyn Polaka and Suzanne Cochrane.....	289
The Time-line as a Device to Enhance Recall in Standardized Research Interviews: A Split Ballot Study Wander van der Vaart.....	301
Using Vignettes in Cognitive Research on Establishment Surveys Rebecca L. Morrison, Kristin Stetler, and Amy E. Anderson.....	319
Pre-printing Effects in Official Statistics: An Experimental Study Anders Holmberg.....	341
Exploring Confidentiality Issues Related to Dependent Interviewing: Preliminary Findings Joanne Pascale and Thomas S. Mayer.....	357
How Good is Good? Comparing Numerical Ratings of Response Options for Two Versions of the Self-Assessed Health Status Question Barbara Foley Wilson, Barbara M. Altman, Karen Whitaker, and Mario Callegaro.....	379
Identifying and Reducing Response Burdens in Internet Business Surveys Gustav Haraldsen.....	393
Book and Software Reviews.....	411

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Dans le numéro de juin 2004, nous avons publié un article rédigé par D.N. Da Silva et Jean D. Opsomer intitulé « Propriétés de l'estimateur à cellules de pondération sous un mécanisme de réponse non paramétrique » (pages 45-55). Nous présentons nos excuses à M. Da Silva pour avoir épelé son nom incorrectement. Celui-ci aurait dû se lire D. Nobrega Da Silva. Veuillez noter également que la version corrigée figure sur le site Web de Statistique Canada.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2004.

- M. Axelsson, *Statistics Sweden*
 D.R. Bellhouse, *University of Western Ontario*
 J.-F. Beaumont, *Statistique Canada*
 M. Avelson, *Statistics Sweden*
 J.M. Brick, *Westat, Inc.*
 D.A. Binder, *Statistique Canada*
 P. Biemer, *Research Triangle Institute*
 D. Cantor, *Westat, Inc.*
 P. Cantwell, *U.S. Bureau of the Census*
 A. Chaudhuri, *Institute of Engineering & Technology Lucknow*
 B.-C. Chen, *U.S. Census Bureau*
 K.R. Copeland, *U.S. Bureau of Labor Statistics*
 J.R. Chromy, *RTI, Inc.*
 P. Dick, *Statistique Canada*
 J.L. Eitinger, *U.S. Bureau of Labor Statistics*
 D. Fogel, *Natural Selection, Inc.*
 O. Frank, *Stockholm University*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistique Canada*
 L. Granquist, *Statistics Sweden*
 B. Graubard, *National Cancer Institute*
 D. Haziza, *Statistique Canada*
 D. Hedeker, *University of Illinois*
 D. Hedin, *Statistics Sweden*
 M.A. Hidington, *Office for National Statistics*
 M. Houbiers, *Statistics Denmark*
 J. Jiang, *University of California at Davis*
 G. Judkins, *Westat, Inc.*
 A. Kennickell, *Federal Research Board*
 P. Kott, *USDA/NAHSS*
 M. Kovacevic, *Statistique Canada*
 J. Kovar, *Statistique Canada*
 M.D. Larsen, *Iowa State University*
 P. Lahiri, *JPSM, University of Maryland*
 R. Lavallee, *Statistique Canada*
 R. Lehtonen, *University of Jyväskylä*
 S. Lohr, *Arizona State University*
 D. Malec, *United States Bureau of the Census*
 H. Maniel, *Statistique Canada*
 D. Marker, *Westat, Inc.*
 S.M. Miller, *U.S. Bureau of Labor Statistics*
 J. Moore, *U.S. Bureau of the Census*
 R. Mukerjee, *Indian Institute of Management*
 G. Nathan, *Hebrew University*
 D. Norris, *Statistique Canada*
 J. Opsomer, *Iowa State University*
 D. Patton, *Statistique Canada*
 D. Pfeffermann, *Hebrew University*
 J.N.K. Rao, *Carleton University*
 T.J. Rao, *Indian Statistical Institute*
 J. Reiter, *Duke University*
 L.-P. Rivest, *Université Laval*
 L. Rizzo, *Westat, Inc.*
 R.A. Rotach, *U.S. Bureau of the Census*
 K. Rust, *Westat, Inc.*
 N. Schenker, *National Opinion Research Center*
 F.J. Scheuren, *National Opinion Research Center*
 I. Schiopu-Kratina, *Statistique Canada*
 N. Shiomu, *Central Bureau of Statistics*
 A.C. Singh, *RTI, Inc.*
 B. Sinha, *ISI*
 M.D. Sinclair, *Mathematica Policy research*
 C.J. Skinner, *University of Southampton*
 P. Smith, *CDC*
 E. Stasny, *Ohio State University*
 D. Steel, *University of Wollongong*
 L. Stokes, *Southern Methodist University*
 C. Swthwarz, *Simon Fraser University*
 A. Tersine, Jr., *United States Bureau of Labor Statistics*
 A. Thierge, *Statistique Canada*
 M. Thompson, *University of Waterloo*
 Y. Tillé, *Université de Neuchâtel*
 C. Tucker, *United States Bureau of Labor Statistics*
 J. van der Brakel, *Central Bureau of Statistics*
 R. Valliant, *JPSM, University of Michigan*
 V. Vehovar, *University of Ljubljana*
 J. Waksberg, *Westat, Inc.*
 J. Wang, *Merck Research Labs, Merck & Co., Inc.*
 F. Wein, *Federal Statistical Office*
 K.M. Wolter, *Iowa State University*
 P. Wong, *Statistique Canada*
 C. Wu, *University of Waterloo*
 Y. You, *Statistique Canada*
 R. Yucel, *University of Massachusetts*
 W. Yung, *Statistique Canada*
 A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 2004: Anne-Marie Fleury, Francine Pilon-Renaud et Roberto Guido (Division de la diffusion), Philippe Laroche (Division du marketing) et François Beaudin (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à Christine Cousineau, Céline Ethier et Denis Lemire de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

- DUNCAN, G.T., KEILNER-MCNUITY, S.A. et STOKES, S.L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Rapport technique, U.S. National Institute of Statistical Sciences.
- FLENNBERG, S.E., MAKOV, U.E. et STEELE, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14, 485-502.
- FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.
- KENNICKELL, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. Dans *Record Linkage Techniques*, 1997, (Eds. W. Alvey et B. Jamerson), Washington, D.C.: National Academy Press, 248-267.
- LITTLE, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.
- LIU, F. et LITTLE, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. Dans *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 2133-2138.
- MENG, X.L. (1994). Multiple-imputation inferences with unconventional sources of input (disc: P558-573). *Statistical Science*, 9, 538-558.
- RAGHUNATHAN, T.E., LEPKOWSKI, J.M., VAN HOEWYK, J. et SOLENNBERGER, P. (2001). Une technique multi-dimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 91-103.
- RAGHUNATHAN, T.E., REITER, J.P. et RUBIN, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.
- REITER, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18, 531-544.
- REITER, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 203-211.
- REITER, J.P. (2004a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*. À paraître.
- REITER, J.P. (2004b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*. À paraître.
- RUBIN, D.B. (1976). Inference and missing data (avec discussion). *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.
- RUBIN, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.
- RUBIN, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57, 3-18.
- SHEN, Z. (2000). *Nested Multiple Imputation*. Thèse de doctorat, Harvard University, Dept. of Statistics.
- WILLENBORG, L. et DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.
- ABOWD, J.M. et WOODCOCK, S.D. (2001). Disclosure limitation in longitudinal linked data. Dans *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, (Eds. P. Doyle, J. Lane, L. Zayatz et J. Theeuwes), Amsterdam: North-Holland, 215-277.
- BARNAARD, J. et RUBIN, D.B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948-955.
- BINDER, D.A., et SUN, W. (1996). Frequency valid multiple imputation for surveys with a complex design. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281-286.
- DALENIUS, T., et REISS, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.

BIBLIOGRAPHIE

Cette étude a été financée par le U.S. Bureau of the Census aux termes d'un contrat conclu avec Datametrics Research. L'auteur remercie R. Little, T. Raghunathan, D. Rubin, L. Zayatz et les examinateurs pour l'inspiration, l'orientation et les commentaires fructueux qu'ils lui ont prodigués.

REMERCIEMENTS

Puisque la variance d'une variable aléatoire centrée moyen est égale à 2 divisé par le nombre de degrés de liberté de la variable, nous concluons que

$$\text{Var} \left(\frac{2f^2}{(m-1)(1+f-g)^2} + \frac{m(r-1)(1+f-g)^2}{2g^2} \right) \quad (31)$$

Si nous combinons (29) et (30), nous voyons que la variance de (23) est approximativement égale à

$$\approx \frac{m(r-1)(1+f-g)^2}{2g^2} \quad (30)$$

$$\text{Var} \left(1 + f - \frac{1}{g} \rho_M \right)$$

méconter certains utilisateurs ou les porter à ne pas avoir confiance dans les données diffusées. On ne sait pas encore très bien comment adapter la méthode – ni, d'ailleurs, de nombreuses autres techniques de contrôle de la divulgation qui modifient les données originales – afin de permettre le calage.

Les données manquantes et le risque de divulgation sont des problèmes importants que doivent résoudre les organismes qui diffusent des données au public. L'approche de l'imputation multiple présentée ici convient pour le traitement simultané de ces deux problèmes, et permet de fournir aux utilisateurs des ensembles de données complets rectangulaires qui peuvent être analysés au moyen de méthodes et de logiciels statistiques standard. La mise en œuvre de cette approche dans des applications réelles pose des défis, mais, comme l'a souligné Rubin (1993) dans sa proposition initiale, les avantages éventuels de cette utilisation de l'imputation multiple sont importants. Le point suivant qu'il conviendra d'étudier est celui de savoir comment la théorie fonctionne en pratique, y compris les comparaisons de l'approche présentée ici à d'autres méthodes de contrôle de la divulgation. Ces comparaisons devraient se concentrer sur des mesures du risque de divulgation obtenues en simulant le comportement d'intrusion et sur des mesures de l'utilité des données pour les paramètres à estimer qui intéressent les utilisateurs, y compris les propriétés des estimations ponctuelles et d'intervalle.

ANNEXE : CALCUL DU NOMBRE APPROXIMATIF DE DEGRÉS DE LIBERTÉ

Les inférences d'après des ensembles de données contenant des imputations multiples pour des données manquantes et le remplacement par des données partiellement synthétiques se font en utilisant une loi de distribution t . Une étape importante consiste à approximer la loi de

$$(23) \quad \left(\frac{u_M^M + (1 + 1/m)B_\infty + B/mr}{v_M^M T^M} \mid d_M \right)$$

sous forme d'une loi du chi-carré à v_M degrés de liberté. On détermine v_M en approximant la moyenne et la variance de la loi χ^2 inverse à la moyenne et à la variance de (23). Alors, $\alpha = (B_\infty + B/r)/B_M$, et soit $\gamma = B/b_M$. $(\alpha^{-1} \mid d_M, B)$ et $(\gamma^{-1} \mid d_M)$ ont une distribution des carrés respectivement. Soit $f = (1 + 1/m)B_M/u_M$, et soit $g = (1/r)b_M^2/u_M$. Nous pouvons écrire (23) sous la forme

$$(24) \quad \frac{T^M}{u_M + (1 + 1/m)B_\infty + B/mr} = \frac{u_M^M(1 + \alpha f - \gamma g)}{\bar{u}^M(1 + f - g)}.$$

Pour approxier les moments, nous devons obtenir une valeur approchée de l'espérance et de la variance de (24). Puisque $\text{Var}(\gamma^{-1} \mid d_M) = 2/(m(r-1))$, le deuxième terme

$$(25) \quad E \left(\frac{1 + \alpha f - \gamma g}{1 + f - g} \mid d_M \right) = E \left(\frac{1 + \alpha f - \gamma g}{1 + f - g} \mid d_M, B \right) = E \left(\frac{1 + \alpha f - \gamma g}{1 + f - g} \mid d_M \right).$$

Nous approximations ces espérances au moyen d'un développement en série de Taylor de premier ordre en α^{-1} et γ^{-1} autour de leurs espérances qui est égal à l'unité. Par conséquent,

$$(26) \quad \approx E \left(\frac{1 + \alpha f - \gamma g}{1 + f - g} \mid d_M \right) \approx 1.$$

Pour la variance, nous utilisons la représentation de la variance conditionnelle

$$(27) \quad E \left(\text{Var} \left(\frac{1 + \alpha f - \gamma g}{1 + f - g} \mid d_M, B \right) \mid d_M \right) + \text{Var} \left(E \left(\frac{1 + \alpha f - \gamma g}{1 + f - g} \mid d_M, B \right) \mid d_M \right).$$

Pour la variance et l'espérance intérieures, nous utilisons un développement en série de Taylor de premier ordre en α^{-1} autour de son espérance. Puisque $\text{Var}(\alpha^{-1} \mid d_M, B) = 2/(m-1)$, l'expression (27) est approximativement égale à

$$(28) \quad E \left(\frac{2(1 + f - g)}{f^2} \mid d_M \right) + \text{Var} \left(\frac{1 + f - g}{1 + f - g} \mid d_M \right).$$

Nous utilisons maintenant des développements en série de Taylor de premier ordre en γ^{-1} autour de son espérance pour déterminer les composantes de (28). Le premier terme

$$(29) \quad \approx \frac{2f^2}{(m-1)(1 + f - g)^2}.$$

statistique peut encore être rehaussée en remplaçant des valeurs pour lesquelles le risque de divulgation est faible (Liu et Little 2002). Ces imputations supplémentaires accroissent la variation des imputations de remplacement et obscurcissent toute information qui pourrait être tirée simplement du fait de savoir quelles données ont été remplacées. Comme pour toute méthode de contrôle de la divulgation (Duncan, Keller-McNulty et Stokes 2001), ces décisions devraient tenir compte du compromis entre le risque de divulgation et l'utilité des données. L'élaboration de lignes directrices sur la sélection des valeurs à remplacer est un sujet de recherche de haute priorité.

Un risque de divulgation persiste dans les données partiellement synthétiques, quelles que soient les valeurs remplacées. En effet, les utilisateurs pourraient se servir des valeurs non modifiées diffusées pour faciliter les tentatives de divulgation, par exemple en les appariant à des bases de données externes, ou être capables d'estimer les valeurs réelles de X_{obs} d'après les données synthétiques avec une précision raisonnable. Par exemple, si la valeur d'une variable de résultat est la même, ou presque, pour toutes les personnes appartenant à un groupe de démographique donné, les modèles d'imputation produiront vraisemblablement cette valeur pour les imputations. Les imputations pourraient alors être obligées de rendre les imputations plus grossières. Un autre exemple est celui où les utilisateurs savent qu'un X_{obs} donné et peuvent donc repérer cet enregistrement enregistré partiellement synthétique.

En ce qui concerne l'utilité des données, le principal défi consiste à spécifier des modèles d'imputation, pour les données manquantes ainsi que pour les données à remplacer, qui donnent des résultats valides. Pour les données manquantes, il est bien connu que des modèles d'imputation peu plausibles risquent de produire des inférences non valides, quoique le problème soit moins prononcé lorsqu'on impute des fractions faibles de données manquantes (Rubin 1987; Meng 1994). Un problème analogue se pose pour les données partiellement synthétiques. Si l'on remplace des fractions importantes de données, par exemple des variables complètes, les analyses portant sur les valeurs remplacées reflètent principalement les hypothèses implícites des modèles d'imputation en ce qui concerne les lois de distribution. Si ces hypothèses sont peu plausibles, les analyses risquent d'être non valides. De nouveau, le problème est moins prononcé si l'on ne remplace que de petites fractions de valeurs, comme on pourrait s'y attendre dans nombre d'applications de l'approche des données partiellement synthétiques.

Certaines caractéristiques des données posent parfois des difficultés particulièrement importantes dans le cas des données partiellement synthétiques. Par exemple, il pourrait

être souhaitable de remplacer les valeurs extrêmes dans le cas de distributions asymétriques, comme les revenus très élevés. L'information au sujet des queues de telles distributions risque d'être limitée, si bien que le tirage de valeurs de remplacement raisonnable tout en protégeant le secret statistique est difficile. Un autre exemple est celui où des valeurs d'imputation sélectionnées au hasard pour des données fortement structurées risquent d'être peu plausibles, comme les combinaisons peu probables d'âge ou d'état matrimonial des membres d'une famille. Ces difficultés, conjuguées aux limitations générales des inférences fondées sur des valeurs imputées, permettent de cerner un domaine de recherche important, celui de l'élaboration et de l'évaluation de méthodes de génération de données partiellement synthétiques, y compris des approches semi-paramétriques et non paramétriques.

Souignons que la construction des modèles de données synthétiques est généralement plus simple que celle des modèles de données manquantes. Les organismes statistiques ne peuvent pas prévoir toutes les utilisations possibles des données diffusées, donc, ne peuvent pas générer des modèles fournissant des résultats valides pour chaque type d'analyse. Un objectif plus modeste et réalisable consiste à permettre aux analystes de faire des inférences valides en utilisant des méthodes et des logiciels standard pour une grande gamme d'analyses type, comme certaines régressions linéaires et logistiques. Les organismes devraient fournir des renseignements permettant aux analystes de décider quelles inférences peuvent être faites à partir des données diffusées. Par exemple, les organismes pourraient donner la description des modèles d'imputation en annexe aux fichiers de données à grande diffusion. Les utilisateurs dont les analyses ne peuvent être réalisées au moyen des données diffusées pourraient être obligés de demander un accès spécial aux données observées. Les organismes doivent aussi fournir de la documentation expliquant comment utiliser les ensembles de données emboîtés. Les règles de combinaison des estimations ponctuelles provenant d'ensembles de données multiples sont suffisamment simples pour être ajoutées aux logiciels statistiques type, comme cela a déjà été fait pour les règles de Rubin (1987) dans SAS, Stata et S-Plus.

Telle que conçue, l'approche d'imputation multiple ne permet pas le calage sur les totaux publiés, ce qui pourrait

$$(u^{(i)} | d_M, B^{(i)}) \sim \left(\bar{u}^{(i)}, << B^{(i)/r} \right) \quad (15)$$

nous supposons que, pour un nombre donné de degrés de liberté, v_M , qu'il convient d'estimer,

$$\left(\frac{\bar{u}_M + (1 + 1/m) B_\infty + B/mr}{v_M^{T_M}} | d_M \right) \sim \chi^2_{v_M} \quad (22)$$

si bien que nous pouvons utiliser une loi t à v_M degrés de liberté pour faire des inférences au sujet de \bar{Q} . Nous obtenons une approximation de v_M en comparant les deux moments de (22) à ceux d'une loi du chi-carré. Les détails montrant que l'expression (9) est une approximation de v_M figurent en annexe.

Les inférences fondées sur les équations (4) à (9) ont des propriétés fréquentistes valides sous certaines conditions. Premièrement, l'analyste doit utiliser des estimateurs valides sous randomisation, q et u . Autrement dit, quand on applique q et u à D pour obtenir q_{obs} et u_{obs} , ($q_{\text{obs}} | X, Y \sim N(\bar{Q}, U)$ et ($u_{\text{obs}} | X, Y \sim (U, << U)$, où la loi pertinente est celle de I . Deuxièmement, les imputations pour les données manquantes doivent être correctes au sens de Rubin (1987, chapitre 4). Essentiellement, il faut pour cela que les inférences d'après les imputations pour les données manquantes soient valides sous randomisation pour q_{obs} et u_{obs} , sous le mécanisme de non-réponse postule.

Troisièmement, les imputations pour les données partiellement synthétiques doivent être synthétiquement correctes au sens de Reiter (2003). Il faut pour cela que les inférences à partir d'imputations de remplacement associées à chaque $D^{(i)}$ soient valides sous randomisation pour $q^{(i)}$ et $u^{(i)}$. En général, il est difficile de vérifier que les imputations pour des données manquantes sont correctes dans le cas d'échantillons complexes (Binder et Sun 1996). Elles pourraient l'être pour certaines analyses, mais non pour d'autres. Par conséquent, certains intervalles de confiance centrés sur des estimateurs sans biais pourraient ne pas avoir de taux de couverture nominal; voir Meng (1994) pour une discussion de cette question. Ces difficultés existent pour l'approche d'imputation multiple utilisée ici et pourraient même être amplifiées à cause de l'imputation supplémentaire de données synthétiques.

5. CONCLUSION

Le recours à l'approche des données partiellement synthétiques pour le contrôle de la divulgation pose de nombreux défis. Par-dessus tout, les organismes statistiques doivent décider quelles valeurs il faut remplacer par des imputations. D'une façon générale, les candidats au remplacement incluent les valeurs de caractéristiques d'identification pour les unités pour lesquelles le risque d'identification est élevé, comme les valeurs d'échantillon uniques et en double, et les valeurs de variables délicates dans les queues des distributions. La protection du secret

où $b^{(i)}$ est défini par (5). Nous supposons ensuite que $B^{(i)} = B$ pour tout i . Cette hypothèse devrait être raisonnable, puisque la variabilité des variances à posteriori a tendance à être d'un ordre plus faible que celle des moyennes à posteriori. En calculant la moyenne sur i , nous obtenons

$$\left((r-1)b^{(i)} B^{(i)} | d_M, B^{(i)} \right) \sim \chi^2_{r-1} \quad (16)$$

où \bar{q}_M est défini par (4) et \bar{u}_M est défini par (7). La loi à posteriori de $(B_\infty | d_M, B)$ est

$$\left(\frac{(m-1)B_M}{B_\infty + B/r} | d_M, B \right) \sim \chi^2_{m-1} \quad (19)$$

où B_M est défini par (6). Enfin, la loi à posteriori de $(B | d_M)$ est

$$\left(\frac{B}{m(r-1)\bar{b}_M} | d_M \right) \sim \chi^2_{m(r-1)} \quad (20)$$

Nous devons intégrer le produit de (11) et (17) en ce qui concerne les lois de distribution dans (19) et (20). Nous pouvons le faire par intégration numérique, mais il est préférable d'obtenir des approximations plus simples pour les utilitateurs.

Pour de grandes valeurs de m et r , nous pouvons remplacer les termes de la variance par leur espérance approximative, c'est-à-dire $B_\infty \approx B_M - B/r$ et $B \approx b_M$. Donc, si m et r sont grands, la loi de distribution à posteriori de \bar{Q} est approximativement :

$$\begin{aligned} (\bar{Q} | d_M) & \sim N(\bar{q}_M, (1 + 1/m)(B_M - \bar{b}_M/r + \bar{u}_M/m)) \\ & = N(\bar{q}_M, (1 + 1/m)B_M - \bar{b}_M/r + \bar{u}_M/m) \\ & = N(\bar{q}_M, T_M). \end{aligned} \quad (21)$$

Si m et r sont de taille moyenne, la loi normale pourrait ne pas être une bonne approximation. Pour dériver une loi d'approximation de référence, nous utilisons les stratégies de Rubin (1987) et de Barnard et Rubin (1999). Autrement dit,

Le comportement de T_M^M et v_M^M dans les cas spéciaux est intéressant. Quand r est très grand, $T_M^M \approx T_m^m$, car $\bar{q}^{(i)} \approx q^{(i)}$, si bien que nous obtenons les résultats de l'analyse de $D^{(i)}$. Si la fraction de valeurs manquantes est faible comparativement à la fraction de valeurs manquantes imputées, b_M^M est faible comparativement à B_M^M , si bien que, de nouveau, $T_M^M \approx T_m^m$. Dans l'un et l'autre cas, v_M^M est approximativement égal à v_m^m , ce qui correspond au nombre de degrés de liberté de Rubin (1987) quand on fait une imputation uniquement pour des données manquantes. Lorsque la fraction de valeurs manquantes est faible comparativement à la fraction de valeurs remplacées, $B_M^M \approx \bar{b}_M^M/r$, si bien que T_M^M est approximativement égal à T_p , avec M ensembles de données diffusés.

4. JUSTIFICATION DES NOUVELLES RÉGLES DE COMBINAISON

La présente section expose une dérivation bayésienne des inférences décrites à la section 3, ainsi que les conditions sous lesquelles ces inférences sont valides du point de vue fréquentiste. Ces résultats s'appuient sur la théorie élaborée par Rubin (1987) et par Reiter (2003). Pour la dérivation bayésienne, nous supposons que l'analyste et l'imputeur utilisent les mêmes modèles.

Soit $D^m = \{d^{(i)} : i = 1, \dots, m\}$ la série complète d'ensembles de données multi-imputés avant que toute valeur observée soit remplacée. Pour chaque $D^{(i)}$, soit $q^{(i)}$ et $u^{(i)}$ la moyenne et la variance de \bar{Q} . À l'instar de Rubin (1987, chapitre 3), soit B_∞ la variance du $q^{(i)}$ obtenu quand $m = \infty$.

Soit $d_M^M = \{d^{(i)} : i = 1, \dots, m\}$ la série complète d'ensembles de données synthétiques diffusés. Pour chaque $d^{(i)}$, soit $q^{(i)}$ la moyenne a posteriori de $q^{(i)}$. Pour chaque l , soit $B^{(l)}$ la variance de $q^{(l)}$ obtenu quand $r = \infty$. Enfin, soit B la moyenne du $B^{(l)}$ obtenu quand $m = \infty$.

En utilisant ces quantités, nous pouvons décomposer la loi a posteriori de $(\bar{Q} | d_M^M)$ comme suit

$$f(\bar{Q} | d_M^M) = \int f(\bar{Q} | d_M^M, D^m, B_\infty, B) f(D^m, B_\infty | d_M^M, B) f(B | d_M^M, D^m, B_\infty, dB). \quad (10)$$

L'intégration est faite sur les lois de distribution des valeurs manquantes dans D et des valeurs remplacées par imputation dans chaque $D^{(i)}$; les valeurs observées, non modifiées demeurent fixes. Nous supposons que les hypothèses asymptotiques bayésiennes sont vérifiées, si bien que les inférences sur données complètes pour \bar{Q} peuvent être fondées sur les lois normales.

Sachant D^m , les données synthétiques sont sans pertinence, si bien que $f(\bar{Q} | d_M^M, D^m, B_\infty, B) = f(\bar{Q} | d_M^M, B_\infty)$. Il s'agit de la loi a posteriori de \bar{Q} pour l'imputation multiple pour les données manquantes, conditionnellement à B_∞ . Comme le montre Rubin (1987), cette loi de distribution a posteriori est approximativement

$$(\bar{Q} | D^m, B_\infty) \sim N(\bar{q}_m, (1 + 1/m)B_\infty + \bar{u}_m) \quad (11)$$

où \bar{q}_m et \bar{u}_m sont définis de la même façon qu'en (1) et (3). Dans le cas de l'imputation multiple pour des données manquantes, nous intégrons (11) sur la loi a posteriori de $(B_\infty | D^m)$. Nous ne le faisons pas ici, puisque nous intégrons sur $(B_\infty | d_M^M)$.

4.2 Évaluation de $f(D^m, B_\infty | d_M^M, B) f(B | d_M^M)$

Puisque la loi de distribution de \bar{Q} donnée par (11) s'appuie uniquement sur \bar{q}_m , \bar{u}_m et B_∞ , il est suffisant, pour $f(D^m, B_\infty | d_M^M, B)$, de déterminer

$$f(\bar{q}_m, \bar{u}_m, B_\infty | d_M^M, B) =$$

$$f(\bar{q}_m, \bar{u}_m | d_M^M, B_\infty, B) f(B_\infty | d_M^M, B).$$

À l'exemple de Reiter (2003), nous commençons par supposer que les imputations pour le remplacement de valeurs sont faites de sorte que, pour tout i , les lois d'échantillonnage pour chaque $q_i^{(i)}$ et $u_i^{(i)}$ soient,

$$(q_i^{(i)} | D^{(i)}, B^{(i)}) \sim N(q^{(i)}, B^{(i)}) \quad (12)$$

$$(u_i^{(i)} | D^{(i)}, B^{(i)}) \sim (u^{(i)}, << B^{(i)}). \quad (13)$$

Ici, la notation $F \sim (G, << H)$ signifie que la variable aléatoire F suit une loi d'espérance G et de variabilité nettement inférieure à H . En réalité, $u_i^{(i)}$ est typiquement centré sur une valeur supérieure à $u^{(i)}$, puisque les données synthétiques introduisent une incertitude due au tirage des valeurs des paramètres. Pour les échantillons de grande taille n , ce biais devrait être minimal. L'hypothèse selon laquelle $E(q_i^{(i)} | D^{(i)}, B^{(i)}) = q^{(i)}$ et l'hypothèse de normalité devraient être raisonnables lorsque les valeurs imputées sont tirées à partir de la loi prédictive a posteriori correcte, $f(Y^{rep} | D^{(i)}, Z^{(i)})$, et que les hypothèses asymptotiques habituelles tiennent.

En supposant que les lois a priori pour tout $q^{(i)}$ et $v^{(i)}$ soient plates, la théorie bayésienne classique sous-entend

$$(q^{(i)} | d_M^M, B^{(i)}) \sim N(\bar{q}^{(i)}, B^{(i)}/r) \quad (14)$$

ajustée en utilisant toutes les données de R^{obs} . En revanche les imputations pour remplacer des valeurs existantes doivent être fondées sur une distribution a posteriori assumée à la condition que les valeurs soient supérieures au seuil. Le tirage de R^{miss} et de R^{rep} à partir de la même loi produira des inférences biaisées.

L'imputation séparée des valeurs de R^{miss} et de R^{rep} crée, en plus de la variabilité d'échantillonnage dans D , deux sources de variabilité dont l'utilisateur doit tenir compte pour obtenir des inférences valides. Ni T^m ni T^p n'estime correctement la variation totale introduite par l'utilisation double de l'imputation multiple. Deux exemples simples illustrent le biais que chacun présente. Supposons que l'on ne doive remplacer qu'une seule valeur, mais que l'on doive imputer des centaines de valeurs manquantes. Intuitivement, T^m devrait donner une bonne approximation de la variance de l'estimateur ponctuel de \bar{Q} , tandis que T^p devrait sous-estimer la variance puisque'il manque un b^m . Par ailleurs, supposons qu'il ne manque qu'une seule valeur, mais que l'on doive en remplacer des centaines. Dans ce cas, T^p devrait donner une bonne approximation de la variance, tandis que T^m devrait surestimer la variance, puisque'il inclut un b^m supplémentaire.

Pour permettre à l'utilisateur d'estimer correctement la variabilité totale, les organismes statistiques peuvent utiliser une méthode en trois étapes pour générer les imputations. Premièrement, l'organisme remplit R^{miss} de valeurs tirées de la loi de distribution a posteriori pour $(X^{miss}|D)$, ce qui donne m ensembles de données complets, $D^{(1)}, \dots, D^{(m)}$. Puis, dans chaque $D^{(l)}$, il sélectionne les unités dont les valeurs doivent être remplacées, c'est-à-dire celles pour lesquelles $Z_j^{(l)} = 1$. Dans de nombreux cas, l'organisme imputera des valeurs pour la même unité dans tous les ensembles $D^{(l)}$ pour éviter de diffuser toute valeur réelle, en est ainsi dans tout l'exposé et nous supposons qu'il sélectionnera l'indice supérieur l de Z . Troisièmement, dans chaque $D^{(l)}$, l'organisme impute les valeurs $R^{rep,l}$ aux unités pour lesquelles $Z_j^{(l)} = 1$, en utilisant la loi de distribution a posteriori pour $(X^{rep,l}|D^{(l)}, Z)$. Ces étapes sont répétées indépendamment $l = 1, \dots, r$ fois pour $l = 1, \dots, m$, afin de générer un total de $M = mr$ ensembles de données, $d^{(l)} = (X^{rep,l}, Y^{(l)}, R^{rep,l}, R^{miss,l}, Z)$, inclut une étiquette individuelle à partir duquel il a quantifié le l de l'ensemble de données $D^{(l)}$ à partir duquel il a été tiré. Ces M ensembles de données sont diffusés au public. La diffusion d'ensembles de données multi-imputés emboîtés de ce genre a également été proposée pour traiter le problème des données manquantes en dehors du contexte du contrôle de la divulgation (Shen 2000; Rubin 2003).

Les analystes peuvent obtenir des inférences valides à partir de ces ensembles de données diffusés en combinant

les inférences faites à partir des ensembles de données individuels. Comme auparavant, soit q l'estimateur de \bar{Q} et soit u l'estimateur de la variance de q utilisés par l'analyste. Nous supposons que ce dernier spécifie q et u en agissant comme si chaque $d^{(l)}$ comprenait, en fait, des données recueillies à partir d'un échantillon aléatoire de (X, Y) fondé sur le plan d'échantillonnage original l . Pour $l = 1, \dots, m$ et $i = 1, \dots, r$, soit $q_i^{(l)}$ et $u_i^{(l)}$, respectivement, les valeurs de q et u dans l'ensemble de données $d^{(l)}$. Les paramètres qui suivent sont nécessaires pour les inférences au sujet de la grandeur scalaire \bar{Q} :

$$(4) \quad \bar{q}^m = \sum_{m=1}^m \sum_{r=1}^r q_i^{(l)} / (mr) = \sum_{m=1}^m \bar{q}^{(l)} / m$$

$$(5) \quad \underline{b}^m = \sum_{m=1}^m \sum_{r=1}^r (q_i^{(l)})^2 / m(r-1) - (\bar{q}^{(l)})^2 / (r-1)$$

$$(6) \quad B^m = \sum_{m=1}^m (\bar{q}^{(l)} - \bar{q}^m)^2 / (m-1)$$

$$(7) \quad \bar{u}^m = \sum_{m=1}^m \sum_{r=1}^r u_i^{(l)} / (mr)$$

Le paramètre $\bar{q}^{(l)}$ est la moyenne des estimations ponctuelles dans chaque groupe d'ensembles de données dont l'indice est l , et \bar{q}^m est la moyenne de ces moyennes sur toutes les valeurs de l . Le paramètre $b^{(l)}$ est la variance des estimations ponctuelles pour chaque groupe d'ensembles de données dont l'indice est l , et b^m est la moyenne de ces variances. Le paramètre B^m est la variance de $\bar{q}^{(l)}$ sur l'ensemble des ensembles de données synthétiques. Enfin, \bar{u}^m est la moyenne des variances estimées de q sur l'ensemble des ensembles de données synthétiques. Sous les conditions décrites à la section 4, l'analyste peut utiliser \bar{q}^m pour estimer \bar{Q} . Une estimation de la variance de \bar{q}^m est donnée par :

$$(8) \quad T^m = (1 + 1/m) B^m - \underline{b}^m / r + \bar{u}^m$$

Quand n , m et r sont grands, on peut fonder les inférences sur la loi normale, $(\bar{Q} - \bar{q}^m) \sim N(0, T^m)$. Si les valeurs de m et r sont moyennes, on peut les fonder sur la loi t , $(\bar{Q} - \bar{q}^m) \sim t_{r^m}(0, T^m)$, avec un nombre de degrés de liberté égal à

$$(9) \quad \nu^m = \left(\frac{(1 + 1/m) B^m}{\bar{q}^m} + \frac{m(r-1) T^m}{(\bar{b}^m/r)^2} \right)^{-1}$$

par exemple d'après les dossiers de recensement ou les bases de sondage. Enfin, nous écrivons les données observées sous la forme $D = (X, Y^{\text{obs}}, I, R)$.

2.1 Imputation multiple pour données manquantes

L'organisme statistique impute les valeurs pour Y^{mis} en les tirant de la loi prédictive bayésienne a posteriori de $(Y^{\text{mis}} | D)$, ou des approximations de cette loi, comme celles de Raghunathan, Lepkowski, Van Hoewyk et Solenberger (2001). Ces tirages sont répétés indépendamment $l = 1, \dots, m$ fois pour obtenir m ensembles de données complètes, $D^{(l)} = (D, Y^{(l)\text{mis}})$. Des imputations multiples plutôt que simples sont utilisées pour permettre aux analystes d'estimer la variabilité due à l'imputation des valeurs manquantes.

Dans chaque ensemble de données imputé $D^{(l)}$, l'analyste estime la quantité de population d'intérêt, Q , au moyen d'un estimateur q , et la variance de q au moyen d'un estimateur u . Nous supposons que l'analyste spécifie q et u en agissant comme si chaque $D^{(l)}$ contenait, en fait, des données recueillies auprès d'un échantillon aléatoire de (X, Y) fondé sur le plan d'échantillonnage original I , autrement dit que q et u sont des estimateurs sur données complètes.

Pour $l = 1, \dots, m$, soit $q^{(l)}$ et $u^{(l)}$, respectivement, les valeurs de q et de u dans l'ensemble de données $D^{(l)}$. Sous obtenir des inférences valides pour la grandeur scalaire Q en combinant les $q^{(l)}$ et $u^{(l)}$. Plus précisément, les quantités qui suivent sont nécessaires pour les inférences :

$$\bar{q}^m = \frac{1}{m} \sum_{l=1}^m q^{(l)} \quad (1)$$

$$b^m = \frac{1}{m} \sum_{l=1}^m (b^{(l)})^2 - \bar{q}^m{}^2 / (m - 1) \quad (2)$$

$$u^m = \frac{1}{m} \sum_{l=1}^m u^{(l)} \quad (3)$$

L'analyste peut alors utiliser \bar{q}^m pour estimer Q et $T^m = (1 + 1/m)b^m + u^m$ pour estimer la variance de \bar{q}^m . Les inférences peuvent être fondées sur des lois t dont le nombre de degrés de liberté est $v^m = (m - 1)(1 + 1/m)b^m$.

2.2 Imputation multiple pour données partiellement synthétiques quand $Y^{\text{inc}} = Y^{\text{obs}}$

En supposant qu'aucune donnée ne manque, c'est-à-dire $Y^{\text{inc}} = Y^{\text{obs}}$, l'organisme construit des ensembles de données partiellement synthétiques en remplaçant certaines valeurs observées par des imputations. Soit $Z_j = 1$, si l'on choisit de remplacer certaines données observées pour l'unité j par des valeurs synthétiques et soit $Z_j = 0$ pour les unités pour

Les inférences d'après des ensembles de données partiellement synthétiques sont fondées sur les quantités définies par les équations (1) à (3). Comme l'a montré Reiter (2003), sous certaines conditions, l'analyste peut utiliser \bar{q}^p pour estimer Q et $T^p = b^p + r^p + u^p$ pour estimer la variance de \bar{q}^p . Les inférences pour la grandeur scalaire Q peuvent être fondées sur des lois t dont le nombre de degrés de liberté est égal à $v^p = (r - 1)(1 + u^p/(b^p/r))^2$.

3. DONNÉES PARTIELLEMENT SYNTHÉTIQUES QUAND $Y^{\text{inc}} \neq Y^{\text{obs}}$

Lorsque certaines données manquent, il semble logique de procéder simultanément à l'imputation des données manquantes et des données partiellement synthétiques. Cependant, l'imputation de Y^{mis} et X^{rep} à partir de la même loi de distribution prédictive a posteriori peut donner des imputations incorrectes. À titre d'exemple, supposons que l'organisme statistique cherche à remplacer toutes les valeurs supérieures à un seuil donné par des imputations. Les imputations pour remplacer les données manquantes peuvent être fondées sur une loi de distribution normale

Les valeurs comprises dans Z peuvent, et il en est variable due à l'imputation.

Les valeurs comprises dans Z peuvent, et il en est souvent ainsi, dépendre des valeurs comprises dans D . Par exemple, l'organisme pourrait ne stimuler des valeurs déli-cates ou des identificateurs que pour les unités de l'échantillon présentant des combinaisons rares d'identificateurs; ou, l'imputeur pourrait remplacer uniquement les valeurs de revenus supérieures à 100 000\$ par des valeurs imputées. Pour éviter d'introduire un biais, les valeurs imputées devraient être tirées de la loi prédictive a posteriori de X pour les unités pour lesquelles $Z_j = 1$. Reiter (2003) illustre les problèmes qui peuvent se poser lorsque les imputations ne sont pas conditionnelles à Z .

Les inférences d'après des ensembles de données partiellement synthétiques sont fondées sur les quantités définies par les équations (1) à (3). Comme l'a montré Reiter (2003), sous certaines conditions, l'analyste peut utiliser \bar{q}^p pour estimer Q et $T^p = b^p + r^p + u^p$ pour estimer la variance de \bar{q}^p . Les inférences pour la grandeur scalaire Q peuvent être fondées sur des lois t dont le nombre de degrés de liberté est égal à $v^p = (r - 1)(1 + u^p/(b^p/r))^2$.

(2004b), l'utilisateur des données peut faire des inférences valides pour toute une gamme de paramètres à estimer au moyen de méthodes statistiques et de logiciels standard applicables à des données complètes. D'autres caractéristiques intéressantes des données entièrement synthétiques sont décrites par Rubin (1993), Little (1993), Fienberg, Makov et Slicke (1998), Raghunathan et coll. (2003) et Reiter (2002, 2004a).

Au moment de la rédaction du présent article, aucun organisme statistique n'avait diffusé des ensembles de données entièrement synthétiques, mais certains avaient adopté une variante de la méthode d'imputation multiple proposée par Little (1993), c'est-à-dire diffuser des ensembles de données comprenant les unités observées au départ, en remplaçant certaines valeurs recueillies, comme les valeurs délicates présentant un risque élevé de divulgation ou les valeurs d'identificateurs clés, par des imputations multiples. Ces ensembles sont appelés ensembles de données *partiellement synthétiques*. Par exemple, le U.S. Federal Reserve Board protège les données de la U.S. Survey of Consumer Finances en remplaçant les valeurs monétaires posant un grand risque de divulgation par des imputations multiples et en diffusant un mélange de ces valeurs imputées et de valeurs non remplacées, recueillies durant l'enquête (Kennickell 1997). Le U.S. Bureau of the Census, ainsi que Abowd et Woodcock (2001) protègent les données des ensembles de données longitudinales couplées en remplaçant toutes les valeurs de certaines variables délicates par des imputations multiples et en retenant les valeurs réelles des autres variables. Liu et Little (2002) présentent un algorithme général, appelé SMiKE, pour simuler des valeurs multiples d'identificateurs clés pour certaines unités.

Toutes ces approches partiellement synthétiques sont séduisantes, parce qu'elles promettent de préserver les avantages principaux des données entièrement synthétiques, c'est-à-dire assurer le respect du secret statistique, tout en permettant aux utilisateurs de faire des inférences sans devoir apprendre à appliquer des méthodes et des logiciels statistiques compliqués, et qu'elles diminuent la sensibilité aux spécifications des modèles d'imputation (Reiter 2003). Des inférences valides à partir d'ensembles de données partiellement synthétiques peuvent être obtenues en suivant les méthodes élaborées par Reiter (2003, 2004b), dont les règles de combinaison des estimations ponctuelles et de variance diffèrent de celles proposées par Rubin (1987) et par Raghunathan et coll. (2003).

Pour décrire l'imputation multiple, nous utilisons la notation de Rubin (1987). Pour une population finie de taille N , soit $I_j = 1$ si l'unité j est sélectionnée dans l'enquête, et $I_j = 0$ autrement, où $j = 1, 2, \dots, N$. Soit $I = (I_1, \dots, I_N)$. Soit R_j un vecteur de dimension $p \times 1$ d'indicateurs de réponse, où $R_{jk} = 1$ si la réponse de l'unité j à la question d'enquête k est enregistrée et $R_{jk} = 0$, autrement. Soit $R = (R_1, \dots, R_N)$. Soit X_j la matrice de dimensions $N \times d$ de données d'enquête pour toutes les unités de la population. Soit $X_{inc} = (X_1^{obs}, \dots, X_N^{inc})$, la matrice $n \times d$ de données d'enquête pour les n unités, avec $I_j = 1$; X_{inc} est la portion de X_{inc} qui est observée, et X_{mis} est la proportion de X_{inc} pour lesquels les données sont manquantes à cause de la non-réponse. Soit X , la matrice de dimensions $N \times d$ de variables de plan de sondage pour l'ensemble des N unités de la population, comme les indicateurs de strate ou de grappe ou les mesures de taille. Nous supposons que ce genre d'information sur le plan de sondage est connu approximativement pour toutes les unités de la population.

2. REVUE DES INFÉRENCES EN PRÉSENCE D'IMPUTATION MULTIPLE

Le but de susciter de futurs travaux de recherche.

Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation

JEROME P. REITER¹

RÉSUMÉ

Plusieurs organismes statistiques utilisent, ou considèrent utiliser, l'imputation multiple pour limiter le risque de divulguer l'identité des répondants ou certains attributs délicats dans les fichiers de données à grande diffusion. Par exemple, ces organismes peuvent diffuser des ensembles de données partiellement synthétiques comportant les unités étudiées originellement, où certaines valeurs recueillies, comme les valeurs délicates posant un risque élevé de divulgation ou les valeurs d'identification clés, sont remplacées par des imputations multiples. Le présent article décrit une approche permettant de générer des ensembles de données partiellement synthétiques multi-imputés pour traiter simultanément le contrôle de la divulgation et les données manquantes. L'idée fondamentale consiste à imputer d'abord les valeurs manquantes pour produire m ensembles de données complètes, puis à remplacer dans chaque ensemble de données complet les valeurs délicates ou permettant l'identification par r valeurs imputées. L'article décrit aussi des méthodes permettant de faire des inférences valides à partir d'ensembles de données multi-imputés de ce genre. De nouvelles règles sont nécessaires pour combiner les estimations ponctuelles et de variances multiples, parce que les deux étapes d'imputation multiple introduisent dans les estimations ponctuelles des sources de variabilité que les méthodes existantes d'obtention d'inférence à partir d'ensembles de données multi-imputés ne mesurent pas correctement. Une loi t de référence appropriée pour l'inférence quand les valeurs de m et r sont moyennes est établie au moyen d'approximations par appartenance de moments et par développement en série de Taylor.

MOTS-CLÉS : Confidentialité; données manquantes; données à grande diffusion; enquête; données synthétiques.

1. INTRODUCTION

Nombre d'organismes statistiques diffusent des micro-données, c'est-à-dire des données sur les unités individuelles, dans les fichiers à grande diffusion. Ces organismes s'efforcent de produire des fichiers à l'abrit des attaques par des utilisateurs de données malintentionnés cherchant à découvrir l'identité ou les attributs des répondants, ii) informationnelles pour une grande gamme d'analyses statistiques et iii) faciles à analyser par les méthodes statistiques standard. Bien faire cela est une tâche difficile. La profiltration des bases de données auxquelles a accès le public et l'amélioration des techniques de couplage d'enregistrements ont rendu le risque de divulgation tellement menaçant que la plupart des organismes statistiques altèrent les micro-données avant leur diffusion. Ainsi, ils recodent globalement les variables, par exemple en diffusant les âges par tranches de cinq ans ou en regroupant les valeurs extrêmes de revenus supérieures à 100 000\$ en une catégorie « 100 000\$ et plus » (Willenborg et de Waal 2001), ils permutent les valeurs d'unités sélectionnées au hasard (Dalenius et Reiss 1982), ou ils ajoutent un bruit aléatoire aux valeurs des variables continues (Fuller 1993). Ces stratégies réduisent inévitablement l'utilité des données diffusées, car elles faussent les résultats de certaines analyses et en rendent d'autres impossibles. Elles compliquent aussi les analyses. Pour analyser correctement des données

perturbées, l'utilisateur doit appliquer les méthodes fondées sur la vraisemblance décrites par Little (1993) ou les modèles d'erreur de mesure décrits par Fuller (1993). Ces méthodes et modèles sont difficiles à utiliser dans le cas de paramètres à estimer non standard et forcent parfois les analystes à apprendre à se servir de nouvelles méthodes statistiques et de programmes logiciels spécialisés. Rubin (1993) a proposé une autre approche en vue de produire des données à grande diffusion, à savoir diffuser des ensembles de données synthétiques multi-imputés. Plus précisément, il propose que les organismes statistiques i) échantillonnent aléatoirement et indépendamment des unités provenant de la base de sondage pour constituer chaque ensemble de données synthétiques, ii) imputent des valeurs inconnues aux unités des échantillons synthétiques en se servant de modèles ajustés d'après les données d'enquête originales et iii) diffusent plusieurs versions de ces ensembles de données au public. Ces ensembles sont appelés ensembles de données *entièrement synthétiques*. La diffusion de données entièrement synthétiques assure le respect du secret statistique, puisque l'identification des unités et de leurs données délicates est quasiment impossible lorsque les valeurs contenues dans l'ensemble de données diffusé ne sont pas des valeurs recueillies réelles. De surcroît, en générant des données synthétiques appropriées et en appliquant les méthodes d'inférence élaborées par Raghunathan, Reiter et Rubin (2003) et par Reiter

- GREEN, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711-732.
- HOEFTING, J.A., MADIGAN, D., RAFTERY, A.E. et VOLINSKY, C. (1999). Bayesian model averaging: a tutorial (avec discussion). *Statistical Science*, 14, 382-417.
- HOLMES, C.C., et MALLICK, B.K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation*, 10, 1217-1233.
- HORNIK, K., STINCHCOMBE, M. et WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359-366.
- KASS, R.E., et VAIDYANATHAN, S. (1992). Approximate Bayesian factor and orthogonal proportions, with applications to testing equality of two binomial proportions. *Journal of the Royal Statistical Society B*, 54, 129-144.
- KURK, A.Y.C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80, 385-392.
- KURK, A.Y.C., et WELSH, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society B*, 63, 277-292.
- LIANG, F., TRUONG, Y.K. et WONG, W.H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statistica Sinica*, 11, 1005-1029.
- LIANG, F., et WONG, W.H. (2001). Real parameter evolutionary Monte Carlo with applications in Bayesian mixture models. *Journal of the American Statistical Association*, 96, 653-666.
- MACKEY, D.J.C. (1992). A practical Bayesian framework for backprop networks. *Neural Computation*, 4, 448-472.
- MADIGAN, D., et RAFTERY, A.E. (1994). Model selection and Occam's window. *Journal of the American Statistical Association*, 89, 1535-1546.
- MARRS, A.D. (1998). An application of reversible-jump MCMC to multivariate spherical Gaussian mixtures. Dans *Advances in Neural Information Processing Systems 10*. San Mateo, CA: Morgan Kaufmann. 577-583.
- MÜLLER, P., et INSUA, D.R. (1998). Issues in Bayesian analysis of neural network models. *Neural Computation*, 10, 749-770.
- NEAL, R.M. (1996). *Bayesian Learning For Neural Networks*. New York: Spnnger-Verlag.
- NETER, J., KUTNER, M.H., NACHTSHEIM, C.J. et WASSERMAN, W. (1996). *Applied Linear Statistical Models* (Quatrième édition). Chicago: Irwin.
- ROBERTS, C.P., et CASELLA, G. (1999). *Monte Carlo Statistical Methods*. New York: Spnnger.
- ROBERTS, G.O. (1996). Markov chain concepts related to sampling algorithms. Dans *Markov Chain Monte Carlo in Practice* (Eds. W.R. Gilks, S. Richardson et D.J. Spiegelhalter). London: Chapman & Hall/CRC. 45-57.
- SÄRNDAAL, C.-B., SWENSSON, B. et WREFTMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- SILVA, P.L.D. (1996). Some asymptotic results on the mean squared error of the regression estimator under simple random sampling without replacement. Southampton: University of Southampton. Center for Survey Data Analysis Technical Report 6-2.
- SILVA, P.L.D., et SKINNER, C. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.
- THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal American Statistical Association*, 94, 635-644.
- TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701-1786.
- VALLIANT, R., DORFMAN, A.H. et ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- WALKER, A.M. (1969). On the asymptotic behaviour of posterior distributions. *Journal Royal Statistics Society, B*, 31, 80-88.
- WEIGEND, A.S., HUBERMAN, B.A. et RUMELHART, D.E. (1990). Predicting the future: A connectionist approach. *Int. J. Neural Syst.*, 1, 193-209.
- WU, C., et SITTER, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal American Statistical Association*, 96, 185-193.

REMERCIEMENTS

Les auteurs remercient Chris Skinner de leur avoir fourni l'ensemble de données de l'essai du recensement, ainsi que les examinateurs anonymes, le rédacteur adjoint et le rédacteur en chef, M.P. Singh, de leur commentaires constructifs qui leur ont permis d'améliorer considérablement l'article.

BIBLIOGRAPHIE

BANKIER, M.D. (1990). Two step generalized least squares estimation. Ottawa: Statistique Canada, Division des méthodes d'enquêtes sociales, Rapport interne.

BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BILLINGSLEY, P. (1986). *Probability and Measure* (Deuxième édition). New York: John Wiley & Sons, Inc.

BREIDT, F.J., et OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

BUNTINE, W.L., et WEIGEND, A.S. (1991). Bayesian back-propagation. *Complex Systems*, 5, 603-643.

CASELLA, G., et BERGER, R.L. (2002). *Statistical Inference* (Deuxième édition). United States: Thompson Learning.

CHAMBERS, R.L., DORFMAN, A.H. et WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

COCHRAN, W.G. (1977). *Sampling techniques* (3ème Ed.). New York: John Wiley & Sons, Inc.

CYBENKO, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303-314.

DENG, L.Y., et WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.

DEVILLE, J.-C., et SÄRNDAHL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DORFMAN, A.H. (1992). Non-parametric regression for estimating totals in finite populations. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 622-625.

DUNSTAN, R., et CHAMBERS, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.

FIRTH, D., et BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society B*, 60, 3-21.

FUNAHASHI, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2, 183-192.

GELMAN, A., et RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (avec discussion). *Statistical Science*, 7, 457-472.

où la première approximation découle du développement en série de Taylor, $\log(1+z) \approx z$, quand z se situe au voisinage de zéro; et la deuxième approximation découle de la loi faible des grands nombres en supposant que n est grand. Notons que l'on fixe souvent v à une valeur élevée, disons un nombre supérieur à 30. Dans le premier exemple donné dans l'article, nous fixons $v = 100$. L'équation (27) implique que le minimum de $h(\theta_v)$ est atteint quand $g(x_j, \theta_v) = g(x_j, \theta_v)$ est vérifiée, c'est-à-dire $g(x_j, \theta_v) = g(x_j, \theta_v)$, où $\theta_v = \arg \min_{\theta} h(\theta_v)$.

Par application de la méthode de Laplace au numérateur de (25) avec $b(\cdot) = g(x_0, \theta_v) \pi(\theta_v | D)$, nous obtenons

$$\sum_{v \in \Omega} P(v) \int g(x_0, \theta_v) \exp(-\tau H(\theta_v)) \pi(\theta_v | v) dv$$
$$\approx \sum_{v \in \Omega} P(v) (2\pi/n)^{m/2} |\sum_v|^{1/2}$$
$$\approx g(x_0) \sum_{v \in \Omega} P(v) (2\pi/n)^{m/2} |\sum_v|^{1/2}$$
$$\exp\{-nh(\theta_v)\} g(x_0, \theta_v) \pi(\theta_v | D)$$
$$\exp\{-nh(\theta_v)\} \pi(\theta_v | D),$$

où la première approximation découle de la formule de Laplace (24), et la deuxième, de l'égalité $\hat{g}(x_j, \theta_v) = g(x_j)$. Ici, nous supposons que le nombre d'unités cachées de chaque v est suffisamment grand pour que $g(\cdot)$ puisse être approximée arbitrairement bien par le réseau dont les poids sont ajustés correctement. Sinon, ce terme prend une valeur faible et est négligeable dans la dernière approximation de (28).

Similairement, en appliquant la méthode de Laplace au dénominateur de (25) avec $b(\cdot) = \pi(\theta_v | D)$, nous obtenons

$$\sum_{v \in \Omega} P(v) \int \exp\{-nh(\theta_v)\} \pi(\theta_v | v) dv$$
$$\approx \sum_{v \in \Omega} P(v) (2\pi/n)^{m/2} |\sum_v|^{1/2} \exp\{-nh(\theta_v)\} \pi(\theta_v | D). (29)$$

Il découle de (28), (29) et l'exactitude d'approximation ($O(n^{-1})$) de la méthode de Laplace que

$$E_{\pi} g(x_0, \theta_{v_1}) \rightarrow g(x_0), \quad (30)$$

quand $n \rightarrow \infty$ et $M \rightarrow \infty$.

Partie (c). Elle découle de (8), de (9), de (30) et du théorème de Slutsky (Casella et Berger 2002). Ce qui achève la preuve.

Tableau 6
Résultats des calculs pour la deuxième étude en simulation avec $v = 100$, $M = 3$ et $\lambda = 4$

Taille	BIAS	EQM	MEQM	Couverture ^a
($\times 10^3$)	($\times 10^5$)	($\times 10^5$)	($\times 10^5$)	(%)
9,20	-0,512	3,36	3,25	92,6

^a Couverture nominale de 95 %.

5. DISCUSSION

Dans le présent article, nous avons étudié l'utilisation de réseaux neuronaux bayésiens pour l'estimation pour popu-

lulation finie. Les résultats numériques montrent que les modèles BNN donnent lieu à une amélioration significative comparativement aux méthodes fondées sur la régression linéaire. Cette amélioration n'est pas due au calcul de la moyenne sur les modèles bayésiens, mais principalement aux modèles BNN. Nous avons également appliqué la méthode de calcul de la moyenne des modèles bayésiens fondés sur la régression linéaire (Liang, Truong et Wong 2001) au même problème et constaté que l'amélioration par rapport aux résultats obtenus par Silva et Skinner (1997) n'est que marginale. Bien que notre implémentation des modèles BNN ne soit pas spécifique aux populations finies, nous ne pensons pas qu'il s'agisse d'un défaut de notre méthode. La généralité de celle-ci donne à penser que son champ d'application est étendu et comprend, par exemple, la régression non linéaire et les séries chronologiques non linéaires (le programme peut être obtenu sur demande auprès du premier auteur). Naturellement, il serait également intéressant d'étudier plus en profondeur la façon d'utiliser l'information connue contenue dans les variables auxiliaires d'une population finie pour l'apprentissage des réseaux neuronaux bayésiens.

ANNEXE

Avant de prouver le théorème 2.1, nous donnons une formule que nous utiliserons dans la preuve.

Formule 5.1 (Méthode de Laplace)

$$\int b(\theta) \exp\{-nh(\theta)\} d\theta = (2\pi/n)^{p/2} \left| \sum^{1/2} \exp\{-nh(\hat{\theta})\} b(\hat{\theta}) \{1 + O(n^{-1})\} \right|, \quad (24)$$

quand $n \rightarrow \infty$, où $b(\cdot)$ est une fonction générale qui ne dépend pas de n , $h(\theta)$ est une fonction d'ordre constant de n et p est la dimension de θ , $\hat{\theta}$ est le maximum de $-h(\theta)$ et $\sum = (D^2 h(\hat{\theta}))^{-1}$ est l'inverse de la matrice hessienne négative évaluée à $\hat{\theta}$.

$$E_{\pi} \left[g(x_0, \theta_v) \middle| \sum_{k=0}^K P(V_k | D) \right]_{2+\delta} = \int \left[g(x_0, \theta_v) \pi(\theta_v | V_k, D) d\theta_v \right]_{2+\delta}$$

Preuve du théorème 2.1
Partie (a) Par définition de l'espérance, $E_{\pi} [g(x_0, \theta_v) | \sum_{k=0}^K P(V_k | D)]$ peut s'écrire

$$\int \left[g(x_0, \theta_v) \pi(\theta_v | V_k, D) d\theta_v \right]_{2+\delta}$$

Il découle de la normalité des lois a posteriori $\pi(\theta_k | V_k, D)$ (Walker 1969) et du fait que la fonction d'activation $\psi(\cdot)$ figurant dans (3) est bornée que l'expression (9) est vérifiée. Walker (1969) a montré que la loi a posteriori est gaussienne en la limite de données d'apprentissage infinies.

Partie (b). Pour une observation $x_0, E_{\pi} g(x_0, \theta_v)$ peut s'écrire

$$E_{\pi} = g(x_0, \theta_v) =$$

$$\sum_{v \in \Omega} P(V) \int g(x_0, \theta_v) \exp\{-nh(\theta_v)\} \pi(\theta_v | V) d\theta_v \quad \sum_{v \in \Omega} P(V) \int \exp\{-nh(\theta_v)\} \pi(\theta_v | V) d\theta_v$$

où

$$\log \pi(\theta_v | V) = -\log \sigma_v^2 - \frac{1}{2} \sum_{i=0}^M I_{a_i} \left(\log \sigma_v^2 + \frac{a_i}{\sigma_v^2} \right) - \frac{1}{2} \sum_{j=0}^M I_{b_j} \left(\log \sigma_v^2 + \frac{b_j}{\sigma_v^2} \right) - \frac{1}{2} \sum_{j=0}^M I_{\lambda_j} \left(\log \sigma_v^2 + \frac{\lambda_j}{\sigma_v^2} \right) - \frac{2}{m} \log(2\pi) + m \log \lambda - \log(m!). \quad (26)$$

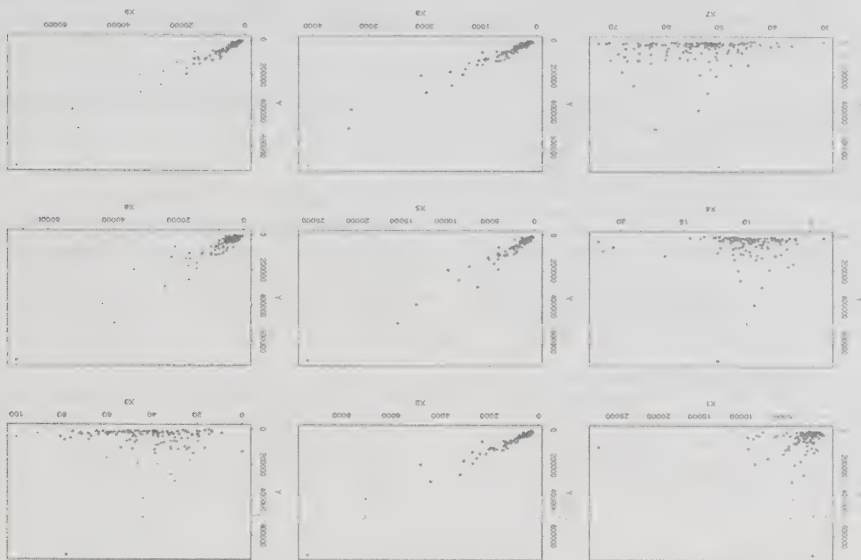
et

$$h(\theta_v) = \frac{1}{n} \log \sigma_v^2 + \frac{2}{v+1} \sum_{i=1}^n \log \left(1 + \frac{V \sigma_v^2}{(y_i - \hat{f}(x_i))^2} \right) \approx \frac{1}{n} \log \sigma_v^2 + \frac{2}{v+1} \sum_{i=1}^n \frac{2}{n} \log \sigma_v^2 + \frac{2}{v+1} \sum_{i=1}^n \frac{V \sigma_v^2}{(y_i - \hat{f}(x_i))^2}$$

$$\approx \frac{1}{v+1} \log \sigma_v^2 + \frac{2V \sigma_v^2}{v+1} E(y_i - \hat{g}(x_i, \theta_v))^2$$

$$= \frac{1}{v+1} \log \sigma_v^2 + \frac{2V \sigma_v^2}{v+1} [E(y_i - g(x_i))^2 + (g(x_i) - \hat{g}(x_i, \theta_v))^2] \quad (27)$$

L'ensemble de données. Pour le deuxième groupe d'expériences, nous avons fixé $v = 100$ et $\lambda = 5$, et fait varier la valeur de M , $M = 1, 2, 3, 4$. Pour le troisième groupe d'expériences, nous avons fixé $v = 100$ et $M = 3$, et fait varier la valeur de λ , $\lambda = 4, 5, 6$. Pour chaque configuration, nous avons exécuté l'algorithme MCESR de la même façon que pour la première étude en simulation. Les résultats des calculs sont résumés au tableau 5. Celui-ci montre que la performance du modèle BNN est assez stable lorsque la configuration des paramètres varie. Il donne aussi à penser que la configuration $v = 100$, $M = 3$ et $\lambda = 4$ serait probablement bonne pour la population simulée en question, si l'on se fonde sur la synthèse des valeurs du



Expériences de validation croisée pour l'échantillon de SMSA. Pour faciliter les comparaisons, les résultats pour la configuration $v = 100$, $M = 3$ et $\lambda = 5$ sont répétés pour les panneaux B et C.

Tableau 5

Expériences de validation croisée pour l'échantillon de SMSA. Pour faciliter les comparaisons, les résultats pour la configuration $v = 100$, $M = 3$ et $\lambda = 5$ sont répétés pour les panneaux B et C.

Expérience	v	M	λ	Taille	BIAS ($\times 10^3$)	EQM ($\times 10^6$)	MEQM ($\times 10^5$)	Couverture ^a
A	50	3	5	10,68	-0,472	4,78	4,19	
	100	3	5	10,74	-0,270	5,04	4,24	92
	∞	3	5	10,74	-0,543	4,76	4,21	92
B	100	1	5	7,29	-0,466	4,63	3,66	89
	100	2	5	9,42	-0,500	4,61	3,91	90
	100	3	5	10,74	-0,527	5,04	4,24	92
	100	4	5	11,66	-0,480	4,74	4,47	91
C	100	3	4	9,56	-0,434	4,68	4,12	92
	100	3	5	10,74	-0,527	5,04	4,24	92
	100	3	6	11,82	-0,455	4,66	4,28	93

Tableau 6
Résultats des calculs pour la deuxième étude en simulation avec
 $v = 100, M = 3$ et $\lambda = 4$

Taille	BAIS	EQM	MEQM	Couverture ^a
($\times 10^3$)	($\times 10^5$)	($\times 10^5$)	($\times 10^5$)	(%)
9,20	-0,512	3,36	3,25	92,6

^a Couverture nominale de 95 %.

5. DISCUSSION

Dans le présent article, nous avons étudié l'utilisation de réseaux neuronaux bayésiens pour l'estimation pour population finie. Les résultats numériques montrent que les modèles BNN donnent lieu à une amélioration significative comparativement aux méthodes fondées sur la régression linéaire. Cette amélioration n'est pas due au calcul de la moyenne sur les modèles bayésiens, mais principalement aux modèles BNN. Nous avons également appliqué la méthode de calcul de la moyenne des modèles bayésiens fondée sur la régression linéaire (Liang, Truong et Wong 2001) au même problème et constaté que l'amélioration par rapport aux résultats obtenus par Silva et Skinner (1997) n'est que marginale. Bien que notre implémentation des modèles BNN ne soit pas spécifique aux populations finies, nous ne pensons pas qu'il s'agisse d'un défaut de notre méthode. La généralité de celle-ci donne à penser que son champ d'application est étendu et comprend, par exemple, la régression non linéaire et les séries chronologiques non linéaires (le programme peut être obtenu sur demande auprès du premier auteur). Naturellement, il serait également intéressant d'étudier plus en profondeur la façon d'utiliser l'information connue contenue dans les variables auxiliaires d'une population finie pour l'apprentissage des réseaux neuronaux bayésiens.

ANNEXE

Avant de prouver le théorème 2.1, nous donnons une formule que nous utiliserons dans la preuve.

Formule 5.1 (Méthode de Laplace)

$$\int b(\theta) \exp\{-nh(\theta)\} d\theta = (2\pi/n)^{p/2} \left| \sum_{i,j} \exp\{-nh(\hat{\theta})\} b(\hat{\theta}) \{1 + O(n^{-1})\} \right| \quad (24)$$

quand $n \rightarrow \infty$, où $b(\cdot)$ est une fonction générale qui ne dépend pas de n , $h(\theta)$ est une fonction d'ordre constant de n et p est la dimension de θ , $\hat{\theta}$ est le maximiseur de $-h(\theta)$ et $\Sigma = (D^2 h(\hat{\theta}))^{-1}$ est l'inverse de la matrice hessienne négative évaluée à $\hat{\theta}$.

Pour la formulation générale de la méthode de Laplace, consulter Kass et Valdyamathan (1992).

Preuve du théorème 2.1

Partie (a) Par définition de l'espérance, $E_{\pi} |g(x_0, \theta_V)|^{2+\delta}$ peut s'écrire

$$E_{\pi} |g(x_0, \theta_V)|^{2+\delta} = \sum_{k=0}^K \pi(V_k | D)$$

$$\left| \int g(x_0, \theta_V) \pi(\theta_V | V_k, D) d\theta_V \right|^{2+\delta}$$

Il découle de la normalité des lois a posteriori $\pi(\theta_k | V_k, D)$ (Walker 1969) et du fait que la fonction d'activation $\psi(\cdot)$ figurant dans (3) est bornée que l'expression (9) est vérifiée. Walker (1969) a montré que la loi a posteriori est gaussienne en la limite de données d'apprentissage infinies.

Partie (b). Pour une observation $x_0, E_{\pi} g(x_0, \theta_V)$ peut s'écrire

$$E_{\pi} = g(x_0, \theta_V) =$$

$$\sum_{V \in \Omega} P(V) \int g(x_0, \theta_V) \exp\{-nh(\theta_V)\} \pi(\theta_V | V) d\theta_V$$
$$\sum_{V \in \Omega} P(V) \exp\{-nh(\theta_V)\} \pi(\theta_V | V) d\theta_V$$

où

$$\log \pi(\theta_V | V) = -\log \sigma^2 - \frac{1}{2} \sum_{j=0}^p I_{\beta_j}^{\alpha_j} \left(\log \sigma^2 + \frac{\alpha_j}{2} \right)$$
$$- \frac{1}{2} \sum_{j=0}^p I_{\beta_j}^{\alpha_j} \left(\log \sigma^2 + \frac{\alpha_j}{2} \right) + \frac{\beta_j}{2}$$
$$- \frac{1}{2} \sum_{j=0}^p I_{\beta_j}^{\alpha_j} \left(\log \sigma^2 + \frac{\alpha_j}{2} \right) + \frac{\beta_j}{2}$$

$$- \frac{2}{m} \log(2\pi) + m \log \lambda - \log(m!), \quad (26)$$

et

$$h(\theta_V) = \frac{1}{n} \log \sigma^2 + \frac{1}{v+1} \sum_{i=1}^n \left(\log \sigma^2 + \frac{2}{v+1} \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma^2} \right)$$
$$\approx \frac{1}{n} \log \sigma^2 + \frac{2}{v+1} \sum_{i=1}^n \frac{1}{\sigma^2} \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma^2}$$
$$\approx \frac{1}{n} \log \sigma^2 + \frac{2}{v+1} \sum_{i=1}^n \frac{1}{\sigma^2} \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma^2}$$

$$\left[E(y_i - g(x_i))^2 + (g(x_i) - \hat{g}(x_i, \theta_V))^2 \right] \quad (27)$$

l'ensemble de données. Pour le deuxième groupe d'expériences, nous avons fixé $v = 100$ et $\lambda = 5$, et fait varier la valeur de M , $M = 1, 2, 3, 4$. Pour le troisième groupe d'expériences, nous avons fixé $v = 100$ et $M = 3$, et fait varier la valeur de λ , $\lambda = 4, 5, 6$. Pour chaque configuration, nous avons exécuté l'algorithme MCESR de la même façon que pour la première étude en simulation. Les résultats des calculs sont résumés au tableau 5. Celui-ci montre que la performance du modèle BNN est assez stable lorsque la configuration des paramètres varie. Il donne aussi à penser que la configuration $v = 100$, $M = 3$ et $\lambda = 4$ serait probablement bonne pour la population simulée en question, si l'on se fonde sur la synthèse des valeurs du

Nous avons poursuivi l'analyse et généré 500 échantillons répétés de taille 70 à partir des 141 enregistrements par la méthode d'échantillonnage aléatoire simple sans remise. Pour chaque répétition, nous avons exécuté l'algorithme MCESR de la même façon que pour la première étude en simulation. Les résultats des calculs sont résumés au tableau 6. Celui-ci montre que le modèle BNN donne aussi de bons résultats pour cette population. Nous avons essayé les autres configurations des paramètres présentées au tableau 5 pour les 500 échantillons répétés. Les résultats des calculs sont tous comparables.

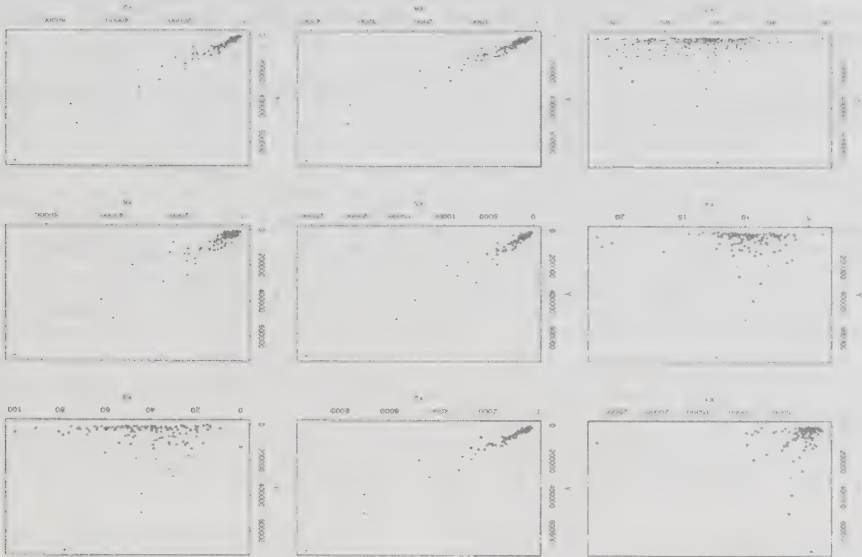


Figure 7 : Diagramme de dispersion de la variable de réponse y en fonction des variables auxiliaires pour la deuxième étude en simulation.

Tableau 5
Expériences de validation croisée pour l'échantillon de SMSA. Pour faciliter les comparaisons, les résultats pour la configuration $v = 100$, $M = 3$ et $\lambda = 5$ sont répétés pour les panneaux B et C.

Expérience	v	M	λ	Taille	BIAS ($\times 10^3$)	EQM ($\times 10^9$)	MEQM ($\times 10^9$)	Couverture ^a (%)
A	50	3	5	5	10,68	-0,472	4,78	91
	100	3	5	5	10,74	-0,270	5,04	92
	∞	3	5	5	10,74	-0,543	4,76	92
B	100	1	5	5	7,29	-0,466	4,63	89
	100	2	5	5	9,42	-0,500	4,61	90
	100	3	5	5	10,74	-0,527	5,04	92
	100	4	5	5	11,66	-0,480	4,74	91
	100	3	4	4	9,56	-0,434	4,68	92
	100	3	5	5	10,74	-0,527	5,04	92
C	100	3	3	6	11,82	-0,455	4,66	93
	100	3	3	5	10,74	-0,527	5,04	92
	100	3	3	4	9,56	-0,434	4,68	92

^a Couverture nominale de 95 %.

Tableau 4
Évaluation de l'influence de v , M et λ sur la taille et la prédictivité du modèle BNN pour le deuxième jeu de variables explicatives. Pour faciliter la comparaison, les résultats de la configuration $v = 100$, $M = 8$ et $\lambda = 5$ ont été répétés dans les panneaux B et C du tableau.

Expérience	v	M	λ	Taille ^a	BIAIS	EQM	MEQM	Couverture ^b
A	50	8	5	14,87	-9,30	394,11	270,09	82,5
	100	8	5	15,06	-5,78	395,25	323,12	86,5
	150	8	5	15,17	-4,38	412,56	346,75	87,1
B	100	6	5	13,90	-5,77	394,79	319,13	86,0
	100	8	5	15,06	-5,78	395,25	323,12	86,5
	100	10	5	16,05	-5,91	396,27	327,86	87,1
C	100	8	4	13,23	-5,62	397,65	323,68	86,4
	100	8	5	15,06	-5,78	395,25	323,12	86,5
	100	8	6	16,76	-5,78	396,45	321,98	86,6

^a Taille = $\sum_{i=1}^{1000} m(A_i) / M / 1000$, où $m(A_i)$ est le nombre de connexions du réseau neuronal A_i .
^b Couverture nominale de 95 %.

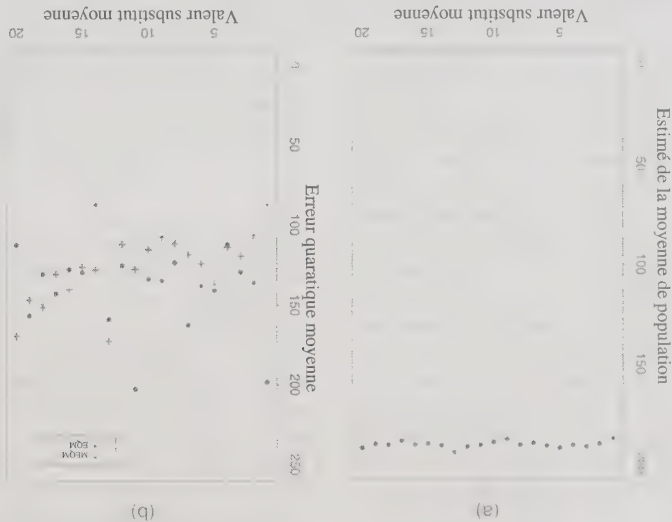
Pour évaluer l'influence de v , M et λ sur la taille et sur la prédictivité du modèle BNN pour le deuxième jeu de variables explicatives, nous avons réalisés les trois mêmes groupes d'expérience que pour le premier jeu de variables explicatives. Les résultats des calculs sont résumés au tableau 4. Le panel A montre de nouveau le compromis entre le BIAIS et l'EQM ou la MEQM pour les modèles BNN selon la valeur choisie de v . Les panels B et C montrent que le BIAIS, l'EQM, la MEQM et la probabilité de couverture sont encore plus stables pour les divers choix des valeurs de M et λ que les mêmes statistiques pour le premier jeu de variables explicatives.

4. DEUXIÈME ÉTUDE EN SIMULATION

Dans la première étude en simulation, nous montrons que le modèle BNN donne de bons résultats pour les ensembles de données avec valeurs aberrantes. Dans la présente étude en simulation, nous montrons que le modèle BNN donne d'encore meilleurs résultats pour les ensembles de données sans valeur aberrante. Dans la présente étude, nous montrons aussi comment une procédure de validation croisée peut être appliquée pour déterminer une configuration des paramètres v , M et λ du modèle BNN. La population simulée comprend les enregistrements de crimes graves recueillis pour 141 grandes régions métropolitaines statistiques normalisées (SMSA) des États-Unis. Une SMSA comprend une ville (ou des villes) dont la taille de population est spécifiée. En général, les données sont celles recueillies pour 1976 et 1977 et sont disponibles dans Neter, Kutner, Nachtsheim et Wasserman (1996). Nous considérons le nombre total de crimes graves en 1977 comme étant la variable d'intérêt (y) et choisissons les neuf variables suivantes comme variables auxiliaires éventuelles.

La figure 7, c'est-à-dire le diagramme de dispersion de y en fonction des neuf variables auxiliaires, donne à penser que le modèle de régression linéaire ne convient peut-être pas pour l'ensemble de données. Il existe une forte relation non linéaire entre y et x_1, x_3, x_4 , ainsi que x_7 . En outre, les variables explicatives x_2, x_5, x_6, x_8 et x_9 sont fortement corrélées. Premièrement, nous démontrons comment on peut appliquer une procédure de validation croisée pour déterminer la configuration des paramètres v , M et λ pour le modèle BNN. Nous avons traité les 70 premiers enregistrements comme s'il s'agissait d'une petite population finie. Nous avons généré 100 échantillons répétés de taille 50 à partir de ces 70 enregistrements par la méthode d'échantillonnage aléatoire simple sans remise, puis nous avons réalisé les expériences suivantes. Pour le premier groupe d'expériences, nous avons fixé $M = 3$ et $\lambda = 5$, et fait varier la valeur de v , $v = 50, 100$ et ∞ , où $v = \infty$ est simplement un indicateur de l'hypothèse de normalité du terme de perturbation. Notons que $M = 3$ donne lieu à un modèle complet de 43 connexions suffisamment grand pour

Figure 6. MOYENNE (panneau (a)), EQM et MEQM (panneau (b)) sachant les valeurs substitués moyennes. Les 1 000 échantillons répétés sont ordonnées sur x_{11} et répartis en 20 groupes de 50 échantillons.



Évaluation de l'influence de v , M et λ sur la taille et sur la prédicivité du modèle BNN pour le premier jeu de variables explicatives. Pour faciliter la comparaison, les résultats de la configuration $v = 100$, $M = 8$ et $\lambda = 5$ sont répétés pour les panneaux B et C.

Expérience	v	M	λ	Taille ^a	BIAIS	EQM	MEQM	Couverture ^b (%)
A	50	8	5	10,53	-6,78	131,78	90,08	82,0
	100	8	5	10,70	-4,91	138,11	127,14	84,8
	150	8	5	10,79	-3,81	156,55	160,28	85,5
B	100	6	5	9,52	-4,90	136,72	122,58	84,1
	100	8	5	10,70	-4,91	138,11	127,14	84,8
	100	10	5	11,83	-5,14	140,13	132,20	86,4
C	100	8	4	9,42	-4,94	138,04	125,99	85,2
	100	8	5	10,70	-4,91	138,11	127,14	84,8
	100	8	6	11,83	-4,92	139,62	128,64	85,7

^a Taille = $\sum_{i=1}^M m(A_i) / M / 1\,000$, où $m(A_i)$ est le nombre de connexions du réseau neuronal A_i .
^b Couverture nominale de 95 %.

Tableau 3

Biais, erreur quadratique moyenne, moyenne des estimations de l'erreur quadratique moyenne et couverture empirique pour diverses stratégies d'estimation de la moyenne de population en utilisant x_1, \dots, x_{10} comme variables auxiliaires. Les autres chiffres BNN sont tirés de Silva et Skinner (1997).

Stratégie d'estimation	BIAIS	EQM	MEQM	Couverture ^a (%)
SM) Moyenne d'échantillon (\bar{y}_i, V_i)	0,25	620,09	619,05	91,8
CN) Réduction du nombre de conditions (\bar{y}_i, V_i)	3,49	562,91	450,36	87,3
RJ) Ridgée	1,05	480,18	472,82	89,4
Fs) Ascendante (\bar{y}_i, V_i)	0,06	468,46	397,99	86,7
Fd) Ascendante (\bar{y}_i, V_d)	-8,12	434,27	338,90	81,7
Fg) Ascendante (\bar{y}_i, V_g)	-7,90	433,71	328,46	81,6
Bs) Meilleure (\bar{y}_i, V_i)	-0,00	466,16	397,59	86,6
BD) Meilleure (\bar{y}_i, V_d)	-7,90	434,54	336,88	81,5
Bg) Meilleure (\bar{y}_i, V_g)	-7,90	433,26	326,05	81,6
FI) Fixe (\bar{y}_i, V_i)	0,45	490,49	461,86	89,0
SS) Sature (\bar{y}_i, V_i)	-0,20	462,71	413,17	86,9
FR) Proc REG (\bar{y}_i, V_i)	-0,07	466,13	399,34	86,4
BNN) $t(100)$	-5,78	395,25	323,12	86,5

^a Couverture nominale de 95 %.

la sous-estimation inévitable de la moyenne de population finie. Il n'existe, franchement, pas grand-chose que l'on puisse faire lorsqu'il existe des valeurs aberrantes dans la population même qu'il n'en existe aucune dans l'échantillon. Aucune méthode statistique fondée uniquement sur l'information provenant de l'échantillon ne peut prédire l'occurrence de valeurs aberrantes dans les données non échantillonnées. Nous pensons que \bar{y}^{BNN} donnera de très bons résultats pour les populations sans valeur aberrante, grâce à la propriété d'approximation universelle des réseaux neuronaux et de la technique de calcul de la moyenne des modèles bayésiens.

Tableau 1

Biais, erreur quadratique moyenne, moyenne des estimations des erreurs quadratiques moyennes et couverture empirique pour diverses stratégies d'estimation de la moyenne de population en utilisant x_1, \dots, x_4 et x_{11} comme variables auxiliaires. Les autres chiffres que ceux pour la stratégie BNN sont tirés de Silva et Skinner (1997).

Stratégie d'estimation		BIAS	EQM	MEQM	Couverture ^a
					(%)
SMA	Moyenne d'échantillon	0,25	620,09	619,05	91,8
(\bar{y}_i, V_i)					
CN	Réduction du nombre de conditions (\bar{y}, V_i)	0,34	507,33	483,63	89,8
RI	Ridge	2,12	304,95	257,07	82,5
(\bar{y}_i, V_i)					
Fed	Ascendante (\bar{y}, V_i)	-1,25	188,08	196,88	82,0
(\bar{y}_i, V_i)					
Ftg	Ascendante (\bar{y}_i, V_i^g)	-1,28	188,38	192,73	81,1
(\bar{y}_i, V_i)					
Bs	Mellieur (\bar{y}_i, V_i)	0,44	236,90	229,49	82,7
(\bar{y}_i, V_i)					
Bd	Mellieur (\bar{y}_i, V_i)	-1,22	190,52	196,84	82,0
(\bar{y}_i, V_i)					
Bg	Mellieur (\bar{y}_i, V_i^g)	-1,24	190,83	192,71	81,1
(\bar{y}_i, V_i)					
FD	Fixe (\bar{y}_i, V_i)	0,29	227,90	241,24	83,3
(\bar{y}_i, V_i)					
Sature	(\bar{y}_i, V_i)	0,30	233,58	242,32	82,5
(\bar{y}_i, V_i)					
FR	Proc REG (\bar{y}_i, V_i)	0,38	235,86	240,26	82,5
(\bar{y}_i, V_i)					
BNN	(100)	-4,91	138,11	127,14	84,8
^a Couverture nominale de 95 %					

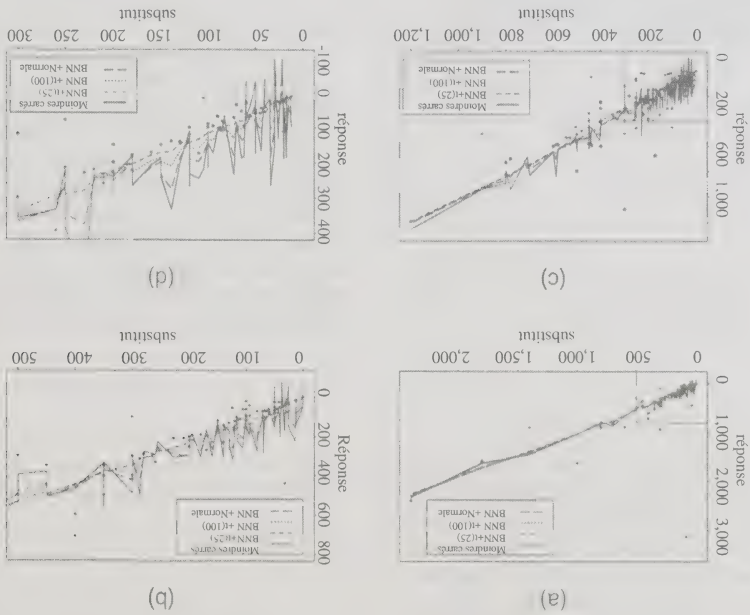


Figure 4. Courbes de réponse ajustée et prédite pour divers modèles. Les courbes sont tracées en fonction de la variable substitut et les valeurs des réponses réelles sont représentées par des points. (a) Courbes de réponse ajustées pour les éléments échantillonnés. (b) Amplification de la région de (a) figurant dans le carré. (c) Courbes de réponse prédites pour les éléments non échantillonnés. (d) Amplification de la région de (c) figurant dans le carré, et par souci de clarté, seul un élément sur quatre est représenté dans la série ordonnée de valeurs substituts triées.

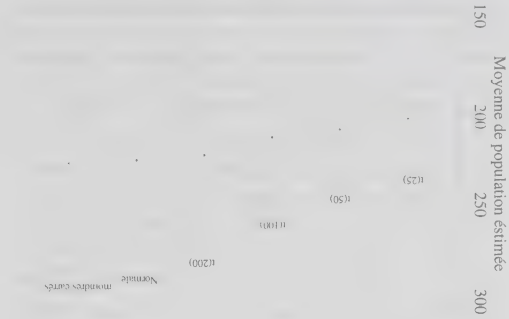


Figure 5. Moyenne de population estimée et intervalle de confiance à 95 % comme pour divers modèles. La ligne en trait pointillé montre la moyenne de population réelle qui est égale à 194,34.

3.4 Résultats numériques sur plusieurs échantillons répétés

Nous avons appliqué les modèles BNN à l'analyse de 1 000 échantillons répétés. Pour chaque échantillon répété du premier jeu de variables explicatives, nous avons fixé $\nu = 100$, $\lambda = 5$ et $M = 8$, ce qui produit 62 connexions pour

stimulée, le MCESR converge ($R < 1,1$) très rapidement, nous avons écarté les 200 premières itérations (100 échantillons utilisés pour le processus de rodage et nous avons utilisé les 600 échantillons restants pour l'inférence. Aux fins de comparaison, nous avons également appliqué le modèle de régression linéaire (1) à cet échantillon répété.

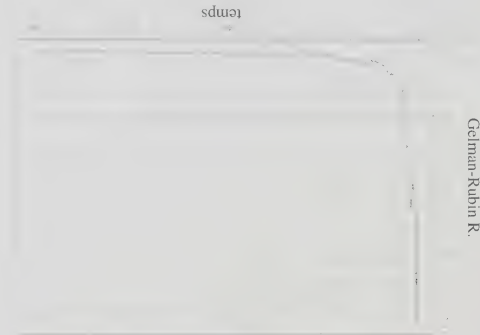


Figure 3. Statistique \hat{R} de Gelman-Rubin. La courbe a été calculée d'après dix exécutions indépendantes de l'algorithme MCESR. Nous supposons que les erreurs aléatoires suivent une loi $t(100)$.

La figure 4 montre les données originales, ainsi que les valeurs ajustées et prédites produites par divers modèles. Les résultats BNN ont tous été obtenus en une seule exécution de l'algorithme MCESR. On voit que le modèle de régression linéaire n'est pas adapté à cette population, car certaines valeurs ajustées et prédites produites par le modèle sont négatives pour cet échantillon répété. En outre, la courbe de réponse ajustée (courbe en trait plein dans les figures 4(a) et 4(b)) est fortement influencée par le 53^e élément et se situe au-dessus de presque les deux tiers des points de données. Un phénomène semblable survient pour la prédiction des valeurs non échantillonnées, voir les figures 4(c) et 4(d)). Par conséquent, la moyenne de population est surestimée (figure 5). Comparativement à ceux du modèle de régression linéaire, les résultats des modèles BNN sont moins influencés par le 53^e élément, particulièrement ceux calculés avec une valeur faible de v . La figure 5 montre qu'à mesure que v diminue, la moyenne de population estimée par les modèles BNN se rapproche de plus en plus de la valeur réelle et que l'intervalle de confiance à 95 % estimé de la moyenne de population devient de plus en plus étroit. Ces résultats indiquent que l'influence du 53^e élément sur ces estimations devient de plus en plus faible à mesure que v augmente. Ce fait n'est pas étonnant, car il est bien connu que l'utilisation d'une distribution de l'erreur à queue lourde rend l'inférence plus robuste.

V_g est une version modifiée d'un estimateur donné par Samdal et coll. (1989) dont les propriétés sont semblables à celles de V_d .

$$V_g = \frac{1-f}{1-f} \sum_{i=1}^d \frac{g_i^2}{g_i^2} \quad (1)$$

La stratégie de sélection du meilleur sous-ensemble (B_s , B_d et B_g) consiste à choisir le sous-ensemble pour lequel l'estimation de l'erreur quadratique moyenne est la plus faible parmi les 2^p sous-ensembles possibles. La stratégie de sélection ascendante (F_s , F_d et F_g) débute en prenant la moyenne d'échantillon comme estimateur, puis en ajoutant la variable qui minimise l'estimation de l'erreur quadratique moyenne et en répétant la procédure jusqu'à ce que l'estimation de l'erreur quadratique moyenne commence à augmenter. Consulter Silva et Skinner (1997) pour des précisions sur l'application des stratégies CN et RL.

3.3 Illustration d'un échantillon répété

Pour comprendre le comportement de \hat{Y}^{BNN} en présence de valeurs aberrantes et le rôle joué par v dans l'inférence robuste, nous nous concentrons sur un exemple particulier. Le jeu de données d'apprentissage comprend les 100 premiers éléments de la population et les variables auxiliaires incluent x_1, \dots, x_d et x_{11} comme premier jeu de variables explicatives. Donc, le 53^e élément a été inclus dans le jeu de données d'apprentissage. Pour chaque configuration des paramètres, nous exécutons l'algorithme MCESR comme suit : les connexions du réseau sont d'abord fixées à certains nombres aléatoires tirés de $N(0, 0.1)$, puis sont mises à jour pour les 1 000 itérations. L'espace paramétrique du modèle complet, autrement dit, la valeur de toutes les variables indicatrices est fixée à 1 dans ces itérations. Après le processus d'initialisation, nous avons exécuté 4 000 itérations de l'algorithme MCESR et nous avons recueilli 800 échantillons à partir de ces itérations au niveau de température le plus faible avec un espace temps égal. La convergence de l'algorithme MCESR peut être diagnostiquée au moyen de la statistique de Gelman-Rubin \hat{R} (Gelman et Rubin 1992) fondée sur des valeurs de \hat{R} calculées d'après dix exécutions indépendantes. Pour chaque échantillon répété de la population

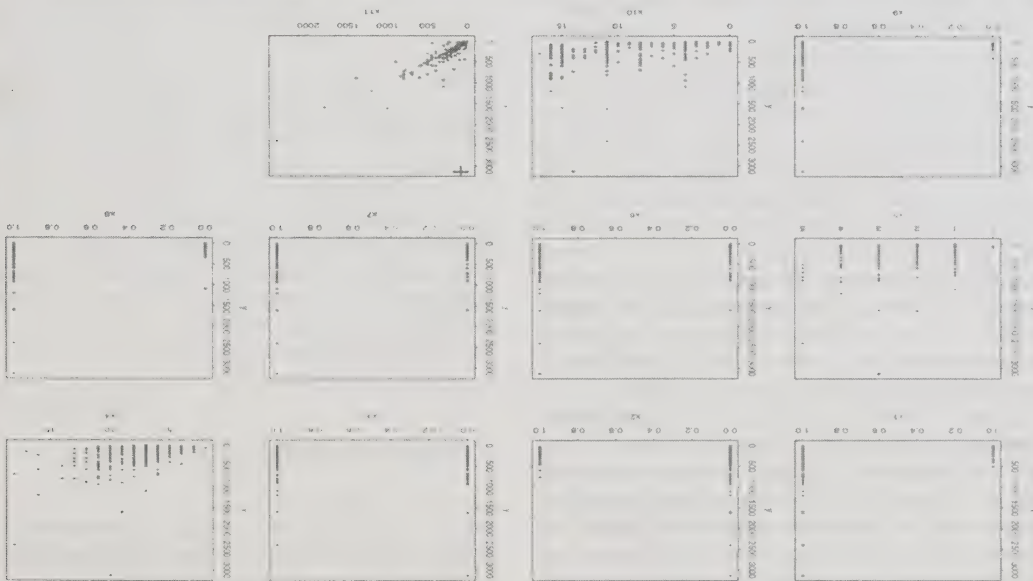


Figure 2. Diagrammes de dispersion de la variable de réponse y en fonction des variables auxiliaires. Dans le graphique de y en fonction de x_{11} , le « + » représente le 53^e élément de la population.

3.2 Revue des stratégies fondées sur la régression

linéaire

Les stratégies fondées sur la régression linéaire qui ont été considérées par Silva et Skinner (1997) sont énumérées ci-après.

SM) Estimateur de la moyenne d'échantillon, sans variables auxiliaires (\bar{y}, V_s) .

FS) Sélection ascendante de variables auxiliaires avec (\bar{y}_r, V_s) .

FD) Sélection ascendante de variables auxiliaires avec (\bar{y}_r, V_d) .

FG) Sélection ascendante de variables auxiliaires avec (\bar{y}_r, V_g) .

BS) Sélection du meilleur sous-ensemble à partir de tous les sous-

ensembles de variables auxiliaires avec (\bar{y}_r, V_s) .

BD) Sélection du meilleur sous-ensemble à partir de tous les sous-

ensembles de variables auxiliaires avec (\bar{y}_r, V_d) .

Bg) Sélection du meilleur sous-ensemble à partir de tous les sous-

ensembles de variables auxiliaires avec (\bar{y}_r, V_g) .

FI) Sous-ensemble fixe de variables auxiliaires avec (\bar{y}_r, V_s) .

SS) Sous-ensemble saturé de variables auxiliaires avec (\bar{y}_r, V_s) .

FR) Sélection ascendante d'un sous-ensemble au moyen de SAS PROC

REG, avec (\bar{y}_r, V_s) .

CN) Procédure de sélection d'un sous-ensemble par réduction du

nombre de conditions avec (\bar{y}_r, V_s) .

RI) Estimateur par la régression ridge proposé par Dunstan et Chambers (1986).

Pour faciliter la description des stratégies susmentionnées, nous définissons la notation qui suit. Soit $U = \{1, \dots, N\}$ une population finie de N éléments distinguables, $D \subset U$ un échantillon répété de n éléments

$$\bar{y}_r = \bar{y} + (X - \bar{X})\beta.$$

V_s, V_d et V_g sont les trois estimateurs de l'erreur quadratique moyenne de \bar{y}_r . V_s , donné par Cochran (1977, page 195), est

$$V_s = \frac{1-f}{n(n-p-1)} \sum_{i \in D} \hat{\epsilon}_i^2,$$

où $\hat{\epsilon}_i = (y_i - \bar{y}) - (x_i - \bar{x})\beta$ et $f = n/N$ est la fraction d'échantillonnage. V_d , qui est généralisé (de $p = 1$ à $p > 1$) à partir d'un estimateur étudié par Deng et Wu (1987) et, en principe, devrait avoir un biais plus faible que V_s (Silva 1996), est

$$V_d = \frac{1-f}{n(n-1)} \sum_{i \in D} u_i \hat{\epsilon}_i^2,$$

L'estimateur par la régression de \bar{y} est $\hat{\beta} = S_{xy}^{-1} S_{xy}$. L'estimateur par les moindres carrés de $\hat{\beta}$ appelle poids g (Särndal, Swensson et Wretman 1989) et $(x_i - \bar{x})(y_i - \bar{y})$, $g_i = 1 + (X - \bar{X})^{-1} S_{xy}^{-1} (x_i - \bar{x})$ ce qu'on réponse, $S_x = n^{-1} \sum_{i \in D} x_i$, la moyenne d'échantillon de la variable de $\bar{y} = n^{-1} \sum_{i \in D} y_i$, le vecteur des moyennes d'échantillon, $\bar{x} = N^{-1} \sum_{i \in U} x_i$, le vecteur des moyennes de population, $\bar{X} = (x_{i1}, \dots, x_{ip})'$ le vecteur de variables remises; $x_i = (x_{i1}, \dots, x_{ip})'$. Soit $\beta = (\beta_1, \dots, \beta_p)'$. Soit

Nous notons que cette estimation est identique en forme à celle donnée par Cochran (1977) pour l'estimateur par la régression linéaire.

3. PREMIÈRE ÉTUDE EN SIMULATION

3.1 Les données

Notre population de simulation comprend 426 enregistrements pour les chefs de ménage faisant partie de l'échantillon qui a répondu au questionnaire (détail) de l'essai de 1988 du recensement de la population de Limeira, dans l'État de São Paulo, au Brésil. Cet essai a été réalisé à titre d'enquête pilote durant les préparatifs du Recensement de la population du Brésil de 1991. Pour une description détaillée de l'essai de recensement, voir Silva et Skinner (1997). À l'instar de ces derniers, nous avons choisi comme variable d'intérêt principale (y) le revenu mensuel total et nous l'avons examinée en même temps que 11 variables auxiliaires possibles, à savoir

x_1	indicateur du sexe du chef du ménage égal masculin;
x_2	indicateur de l'âge du chef du ménage plus petit ou inférieur à 35;
x_3	indicateur de l'âge du chef du ménage supérieur à 35 et inférieur ou égal à 55;
x_4	nombre total de pièces dans le logement;
x_5	nombre total de salles de bain dans le logement;
x_6	indicateur de propriétaire du logement;
x_7	indicateur que le type de logement du ménage est une maison;
x_8	indicateur de propriété d'au moins une automobile par le ménage;
x_9	indicateur de propriété d'un véhicule couleux par le ménage;
x_{10}	nombre d'années d'études du chef de ménage;
x_{11}	variable substitut du revenu mensuel total du chef du ménage.

La figure 2, qui donne les diagrammes de dispersion de y en fonction de 11 variables auxiliaires, montre qu'un modèle de régression linéaire ne convient pas pour représenter des données. Bien que y et x_{11} présentent une forte corrélation linéaire, les graphiques de dispersion de y en fonction de certaines autres variables auxiliaires, disons x_4 , x_5 et x_{10} , donnent à penser que ces relations ne peuvent pas être modélisées convenablement par une régression linéaire. En outre, si nous modélisons les données au moyen d'une régression linéaire, la valeur aberrante, c'est-à-dire le 53^e élément, pourrait avoir une grande influence sur l'ajustement et sur la prédiction du modèle. Plus précisément, si cet élément est inclus dans le jeu de données d'apprentissage, la courbe de réponse ajustée présentera une dérive vers le haut comparativement à la courbe réelle et, par conséquent, la moyenne de population

finie sera surestimée; si l'élément n'est pas inclus dans le jeu de données d'apprentissage, la prédiction se fera comme s'il n'existait pas de valeur aberrante et, par conséquent, la moyenne de population finie sera sous-estimée. La présence d'un élément fortement influent représente aussi un grand défi pour les modèles BNN et d'autres stratégies d'analyse des données.

À l'exemple de Silva et Skinner (1997), nous avons constitué deux jeux distincts de variables auxiliaires pour les simulations. Le premier contient x_1, \dots, x_4 et x_{11} , qui inclut la variable substitutive x_{11} et a un pouvoir explicatif raisonnable pour la prédiction de y . Le deuxième, qui contient x_1, \dots, x_{10} , a un pouvoir explicatif plus faible que le premier à cause de l'exclusion de x_{11} . Donc, ces deux jeux illustrent les propriétés prédictives des modèles BNN avec variables auxiliaires fortes et faibles, respectivement. Comme Silva et Skinner (1997), nous avons sélectionné à partir de cette population simulée 1 000 échantillons répétés de taille 100 par échantillonnage aléatoire simple sans remise. Les calculs qui suivent ont été exécutés sur les 1 000 répliques.

Pour chaque réplique, disons k , les données ont été analysées au moyen de modèles BNN et de diverses stratégies fondées sur la régression linéaire (passées en revue plus loin). Pour toute stratégie, l'estimation de la moyenne de population et celle de son erreur quadratique moyenne pour la réplique k sont représentées par $\bar{y}(k)$ et $V(\bar{y}(k))$, respectivement. Pour résumer les résultats, nous avons calculé la moyenne (MOYENNE), le biais (BIAS), l'erreur quadratique moyenne (EQM) et la moyenne des estimations de l'erreur quadratique moyenne (MEQM) pour l'ensemble de 1 000 répliques, données respectivement par

$$\begin{aligned} \text{MOYENNE} &= \sum_{k=1}^K \bar{y}(k) / S; \\ \text{BIAS} &= \text{MOYENNE} - \bar{y}; \\ \text{EQM} &= \sum_{k=1}^K [\bar{y}(k) - \bar{y}]^2 / S; \\ \text{MEQM} &= \sum_{k=1}^K V(\bar{y}(k)) / S, \end{aligned}$$

où S est le nombre total d'échantillons répétés considérés et $\bar{y} = 194,34$ pour la population simulée. Nous avons également calculé les taux de couverture empiriques des intervalles de confiance à 95 % fondés sur la théorie normale asymptotique pour chaque stratégie et nous les présentons, exprimés en pourcentage, dans la dernière colonne des tableaux 1 et 3.

les propriétés de l'erreur quadratique moyenne décrite plus loin pour tout plan de sondage ignorable. Deuxièmement, cet estimateur est identique à celui proposé par Dorfman (1992), sauf que le BNN est remplacé par une régression par la méthode du moyau. Troisièmement, cet estimateur peut être utilisé pour estimer la moyenne d'une population finie à condition que chaque élément non échantillonné ait la même loi que l'échantillon D .

L'exactitude d'une estimation peut être mesurée par son erreur quadratique moyenne $E(\bar{y}^{\text{BNN}} - \bar{Y})^2$, où \bar{Y} représente la moyenne de population réelle. Pour estimer $E(\bar{y}^{\text{BNN}} - \bar{Y})^2$, nous considérons d'abord

$$E \left[\bar{y}^{\text{BNN},k} - \bar{Y}^2 \mid D, X_N^{n+1} \right] = E \left[\frac{1}{M} \sum_{m=1}^M \sum_{l=n+1}^L \hat{g}(x_l, \theta_{\nu_l}) - \frac{1}{N} \sum_{l=1}^N (\hat{g}(x_l, \nu_l) + \epsilon_l) \right] \left| D, X_N^{n+1} \right] \\ = \left[\frac{M(N-n)}{1} \sum_{m=1}^M \sum_{l=n+1}^L \hat{g}(x_l, \theta_{\nu_l}) - \frac{N-n}{1} \sum_{l=1}^L \hat{g}(x_l, \theta_{\nu_l}) - \frac{N-n}{1} \sum_{l=1}^L \hat{g}(x_l, \nu_l) \right] \left| D, X_N^{n+1} \right] \\ + \frac{N-n}{2} \text{var}(\epsilon_l) + \frac{N-n}{2} \text{var}(\epsilon_l)$$

$$= \frac{(N-n)^2}{N^2} E \left[\left\{ \frac{M(N-n)}{1} \sum_{m=1}^M \sum_{l=n+1}^L \hat{g}(x_l, \theta_{\nu_l}) - \frac{N-n}{1} \sum_{l=1}^L \hat{g}(x_l, \theta_{\nu_l}) - \frac{N-n}{1} \sum_{l=1}^L \hat{g}(x_l, \nu_l) \right\}^2 \right] \left| D, X_N^{n+1} \right] \\ + \frac{N-n}{2} \text{var}(\epsilon_l) + \frac{N-n}{2} \text{var}(\epsilon_l)$$

$$= \frac{(N-n)^2}{N^2} E \left[\left\{ -E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) + E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) - E(\bar{y}^{\text{BNN}}) \right\}^2 \right] \left| D, X_N^{n+1} \right]$$

$$\approx \frac{\tau_D^2}{2} + (1-f)^2 \{E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) - E(\bar{y}^{\text{BNN}})\}^2 \\ + \frac{N^2 \text{var}(\epsilon_l)}{N-n}$$

$$+ \frac{1-f}{N} \text{var}(\epsilon_l), \quad (16)$$

où $X_N^{n+1} = (x_{n+1}^*, \dots, x_N^*)$ représente l'ensemble de vecteurs auxiliaires des éléments non échantillonnés; \bar{y}^{BNN} représente la valeur moyenne de la variable étudiée des éléments non échantillonnés, et

$$E(\bar{y}^{\text{BNN}}) = \frac{1}{N-n} \sum_{l=n+1}^L \hat{g}(x_l, \nu_l).$$

La dernière approximation de (16) découle de (15), autrement dit, si M est grand,

$$E \left\{ \frac{1}{MN} \sum_{m=1}^M \sum_{l=n+1}^L \hat{g}(x_l, \theta_{\nu_l}) - (1-f)E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) \right\}^2 \approx \frac{M}{\tau_D^2}.$$

pour une constante positive τ_D^2 . Le terme $E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) - E(\bar{y}^{\text{BNN}})$ est le biais d'échantillonnage de D . caractéristique aléatoire ou au biais d'échantillonnage de D .

Parlant de (16), nous avons

$$E(\bar{y}^{\text{BNN}} - \bar{Y})^2 \approx \frac{M}{\tau_D^2} + (1-f)^2 E \{ E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) - E(\bar{y}^{\text{BNN}}) \}^2 + \frac{N}{1-f} \text{var}(\epsilon_l). \quad (17)$$

La quantité τ_D^2 peut être estimée par la méthode des moennes de Markov pour $M = rs$ itérations, où s est la taille du lot que l'on suppose suffisamment grande pour que

$$\bar{y}^{\text{BNN},k} = \bar{y} + \frac{1}{s} \sum_{l=1}^s \sum_{i=k-1+s+1}^{i=n+1} \hat{g}(x_i, \theta_{\nu_i}),$$

soit approximativement indépendamment $N(f\bar{y} + (1-f)E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}), \tau_D^2/s)$. Par conséquent, τ_D^2 peut être approximée par

$$\tau_D^2 = \frac{r-1}{s} \sum_{k=1}^s (\bar{y}^{\text{BNN},k} - \bar{y}^{\text{BNN}})^2, \quad (18)$$

que l'on peut introduire par substitution dans (17) pour remplacer $E\tau_D^2$. Sous l'hypothèse de $\epsilon_l \sim t(v)$, l'estimateur par la MMB $\text{var}(\epsilon_l)$ est

$$\text{var}(\epsilon_l) = \frac{1}{v} \frac{M}{1} \sum_{i=1}^M \hat{g}_i^2. \quad (19)$$

Sous l'hypothèse que la population est constituée de copies exactes des données d'apprentissage, nous avons $E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) \approx \bar{y} - \bar{y}$, où \bar{y} représente la moyenne d'échantillon ajustée, et

$$E(\bar{y} - \bar{y})^2 = E \left\{ \frac{1}{n} \sum_{i=1}^n \epsilon_i \right\}^2 = \frac{1}{N} \text{var}(\epsilon_l), \quad (20)$$

où $\epsilon_i = \sum_{j=1}^M \hat{g}(x_i, \theta_{\nu_i})/M - y_i$ est le résidu du i^{e} élément de D , et les ϵ_i sont, par hypothèse, iid et $E(\epsilon_i) = 0$. Sous le modèle réel, nous avons $\text{var}(\epsilon_i) \approx \text{var}(\epsilon_l)$. Donc, nous proposons d'estimer $E\{E(\bar{y}^{\text{BNN}} \mid D, X_N^{n+1}) - E(\bar{y}^{\text{BNN}})\}^2$ par

$$\text{Biais}^2 = \frac{1}{1} \text{var}(\epsilon_l). \quad (21)$$

Sommairement, $E(\bar{y}^{\text{BNN}} - \bar{Y})^2$ peut être estimée par

$$\hat{E}(\bar{y}^{\text{BNN}} - \bar{Y})^2 = \frac{\tau_D^2}{2} + (1-f)^2 \text{Biais}^2 + \frac{N}{1-f} \text{var}(\epsilon_l) = \frac{\tau_D^2}{2} + \frac{M}{1-f} \text{var}(\epsilon_l). \quad (22)$$

Quand $M \rightarrow \infty$, nous avons

$$\hat{E}(\bar{y}^{\text{BNN}} - \bar{Y})^2 = \frac{1}{1-f} \text{var}(\epsilon_l). \quad (23)$$

Supposons que les échantillons $(\theta_1, \Delta_1), \dots, (\theta_M, \Delta_M)$ soient tirés de la loi a posteriori $\pi(\theta | D)$ par un algorithme MCMC. Alors, nous pouvons estimer $p(\theta_{\Delta_i})$ par

$$(6) \quad \hat{p}(\theta_{\Delta_i}) = \frac{1}{M} \sum_{i=1}^M p(\theta_{\Delta_i}),$$

où $\theta_{\Delta_i} = (\theta_i, \Delta_i)$. En appliquant la théorie standard des chaînes de Markov (Tierney 1994; Roberts et Casella 1999), sous des conditions de régularité, nous avons les résultats qui suivent. Si $E_{\pi} |p(\theta_{\Delta_i})| < \infty$, alors

$$(7) \quad \frac{1}{M} \sum_{i=1}^M p(\theta_{\Delta_i}) \rightarrow E_{\pi} p(\theta_{\Delta_i}), \text{ a.s.,}$$

quand $M \rightarrow \infty$. En outre, si $E_{\pi} |p(\theta_{\Delta_i})|^{2+\delta} < \infty$ pour un certain $\delta > 0$, alors

$$(8) \quad M^{1/2} \left\{ \frac{1}{M} \sum_{i=1}^M p(\theta_{\Delta_i}) - E_{\pi} p(\theta_{\Delta_i}) \right\} \rightarrow N(0, \tau_z^2),$$

pour une constante positive τ_z^2 quand $M \rightarrow \infty$, et il y a convergence en loi.

Similairement à (7) et (8), nous avons, pour les modèles BNN, le théorème qui suit dont la preuve est présentée en annexe.

Théorème 2.1 Soit $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ un échantillon aléatoire simple tiré à partir d'une population qui peut être modélisée par le modèle (2). Soit $(\theta_1, \Delta_1), \dots, (\theta_M, \Delta_M)$ l'échantillon tiré à partir de la loi a posteriori $\pi(\theta_{\Delta_i} | D)$, donnée par (4), par une méthode MCMC. Alors, pour tout x_0 tiré de la même loi avec les observations D , nous avons

$$(9) \quad E_{\pi} [g(x_0, \theta_{\Delta_i}) | \tau_z^2 < \infty], \text{ pour un } \delta > 0, \text{ quand } n \rightarrow \infty,$$

$$(10) \quad \frac{1}{M} \sum_{i=1}^M g(x_0, \theta_{\Delta_i}) \rightarrow g(x_0), \text{ a.s.,}$$

$$(11) \quad M^{1/2} \left[\frac{1}{M} \sum_{i=1}^M g(x_0, \theta_{\Delta_i}) - g(x_0) \right] \rightarrow N(0, \tau_z^2),$$

quand $n \rightarrow \infty$ et $M \rightarrow \infty$.

pour une constante positive τ_z^2 quand $n \rightarrow \infty$ et $M \rightarrow \infty$, et il y a convergence en loi.

Pour montrer certaines propriétés des moments de $1/M \sum_{i=1}^M g(x_0, \theta_{\Delta_i})$, nous avons besoin du théorème suivant (Billingsley 1986, page 348, Corollaire).

Théorème 2.2 Soit r un nombre entier positif. Si $X_M \rightarrow X$ en loi et $\sup_m E|X|^{r+\delta} < \infty$, où $\delta > 0$, alors $E|X|^r < \infty$ et $EX_M^r \rightarrow EX^r$.

Il découle de (9), de (11) et du théorème 2.2 que

$$(12) \quad ME \left[\frac{1}{M} \sum_{i=1}^M g(x_0, \theta_{\Delta_i}) - g(x_0) \right] \xrightarrow{2} \tau_z^2,$$

quand $n \rightarrow \infty$ et $M \rightarrow \infty$. Cela implique que

$$(13) \quad E \left[\frac{1}{M} \sum_{i=1}^M g(x_0, \theta_{\Delta_i}) - g(x_0) \right]^2 = \frac{M}{\tau_z^2} + o \left(\frac{1}{M} \right)$$

tient quand n et M sont tous deux grands. Notons que nous avons montré que (11) et (13) tiennent quand la taille de l'échantillon $n \rightarrow \infty$. Dans le contexte de population finie, surtout une petite population finie, des expressions plus précises de (11) et de (13) seraient

$$(14) \quad M^{1/2} \left[\frac{1}{M} \sum_{i=1}^M g(x_0, \theta_{\Delta_i}) - E(y_0 | D, x_0) \right] \rightarrow N(0, \tau_z^2),$$

$$\text{et} \quad E \left[\frac{1}{M} \sum_{i=1}^M g(x_0, \theta_{\Delta_i}) - E(y_0 | D, x_0) \right]^2 = \frac{M}{\tau_z^2} + o \left(\frac{1}{M} \right), \quad (15)$$

où $E(y_0 | D, x_0)$ représente la prédiction de y_0 qui est la variable d'intérêt correspondant à x_0 . Les équations (14) et (15) tiennent compte du biais éventuel de l'échantillon D . Dans le cas où la population est constituée de nombreuses copies exactes de l'échantillon D , $E(y_0 | D, x_0) = g(x_0)$ tient et les équations (14) et (15) se réduisent à (11) et (13), respectivement.

2.2.2 Estimateurs par la moyenne des modèles bayésiens (MMB) dans le cas des populations finies

Considérons une population finie de N éléments distinguables. Au i^{e} élément sont associées la variable d'intérêt y_i et les variables auxiliaires x_i . Les valeurs x_1, \dots, x_N sont connues pour toute la population, tandis que y_i est connue uniquement si la i^{e} unité est sélectionnée dans l'échantillon. Supposons qu'un échantillon aléatoire simple $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ a été tiré à partir de la population finie, qu'un modèle BNN a été conçu pour l'échantillon et que $(\theta_1, \Delta_1), \dots, (\theta_M, \Delta_M)$ ont été tirés de la loi a posteriori du modèle BNN. Alors, l'estimateur par la MMB de la moyenne de la population finie est

$$\bar{y}_{\text{BNN}} = f \bar{y} + \frac{1}{M} \sum_{i=1}^M \sum_{j=n+1}^N g(x_j, \theta_{\Delta_i}),$$

où \bar{y} est la moyenne d'échantillon de y_1, \dots, y_n , et $f = n/N$ est la fraction d'échantillonnage. Nos commentaires au sujet de cet estimateur sont les suivants. Premièrement, \bar{y}_{BNN} est un estimateur fondé sur un modèle, si bien que toutes les inférences pour les y_i sont faites par rapport au modèle et non par rapport au plan de sondage. Aussi longtemps que le modèle tient, l'estimateur BNN aura

de tous les modèles possibles avec $3 \leq m \leq U$. Nous fixons la valeur minimale de m à trois en nous fondant sur les opinions suivantes : les réseaux neuronaux sont habituellement utilisés pour résoudre des problèmes complexes et trois s'est avéré être un nombre suffisamment faible en tant que taille limite de réseau. Dans les lois a priori sus-mentionnées, $\sigma_a^2, \sigma_b^2, \sigma_y^2$ et λ sont des hyperparamètres qui doivent être spécifiés par l'utilisateur (voir la discussion plus loin). De surcroît, nous supposons que ces lois a priori sont indépendantes a priori. Donc, nous avons le log de la probabilité a posteriori (jusqu'à une constante additive) suivant

$$\log \pi(\theta | D) = \text{Constante} - \left(\frac{2}{n} + 1 \right) \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n \log \sigma^2$$

$$\log \left(1 + \frac{(\mathbf{y}_i - \hat{\delta}(\mathbf{x}_i; \theta^V))^2}{\sigma^2} \right)$$

$$- \frac{1}{2} \sum_{i=1}^n \log \sigma_a^2 - \frac{\sigma_a^2}{2} \left(\frac{1}{M} \sum_{j=1}^M I_{ij} \delta_j \right)$$

$$\left(\log \sigma_b^2 + \frac{\sigma_b^2}{2} \right)$$

$$- \frac{1}{2} \sum_{j=1}^M \sum_{i=1}^n I_{ij} \log \sigma_y^2 + \frac{\sigma_y^2}{2} \left(\frac{1}{M} \sum_{j=1}^M I_{ij} \right)$$

$$+ m \log \lambda - \log(m!). \quad (4)$$

Notre modèle BNN diffère de ceux décrits dans la littérature à deux égards importants. Premièrement, les variables d'entrée de notre modèle sont sélectionnées automatiquement par échantillonnage à partir de la loi a posteriori conjointe de la structure du réseau et des poids. Deuxièmement, la structure de notre modèle BNN est habituellement creuse et sa performance dépend moins de la spécification initiale des profils d'entrée et du nombre d'unités cachées. Le modèle est conçu en ce sens que seul un petit nombre de connexions sont actives dans le réseau. Donc, notre modèle de BNN permet d'éviter le problème du surajustement de façon plus naturelle.

En ce qui concerne la préparation des données et l'établissement des hyperparamètres, nos suggestions sont les suivantes. Pour éviter certains poids qui deviennent très grands ou très petits (en valeur absolue) durant l'apprentissage pour pouvoir utiliser diverses échelles de variables d'entrée et de sortie, nous recommandons de normaliser toutes les variables d'entrée et de sortie avant de procéder à l'alimentation des réseaux. Dans tous les exemples présentés ici, les données sont normalisées au moyen de $(y_i - \bar{y})/S_y$ et S_y représentent la moyenne et l'écart-type des données d'apprentissage, respectivement. Considérant qu'un réseau dont les poids varient fortement a habituellement de mauvaises propriétés de généralisation, nous proposons de choisir σ_a^2, σ_b^2 et σ_y^2 pour les valeurs

modérées de façon à pénaliser une forte variation des poids. Par exemple, nous fixons $\sigma_a^2 = \sigma_b^2 = \sigma_y^2 = 5$ pour tous les exemples donnés ici. Cette consigne devrait également être bonne pour les autres problèmes. La valeur de λ reflète notre pensée quant à la taille de réseau nécessaire pour les données étudiées. Ici, comme le suggèrent Weigend, Huberman et Rumelhart (1990), nous choisissons λ de façon à ce que le nombre de poids de connexion soit égal à environ le dixième de la taille de l'échantillon d'apprentissage dans une simulation. Dans une simulation, nous avons évalué l'influence de λ sur la taille et le pouvoir de prédiction du modèle BNN. Les résultats numériques donnent à penser que les modèles BNN ont une prédictivité assez robuste à la variation de λ , quoique leur taille augmente lentement à mesure que λ augmente.

Pour tirer un échantillon à partir de la loi a posteriori (4), nous développons un algorithme de Monte Carlo, appelé algorithme de Monte Carlo échantillonné à saut réversible (MCESR). Cet algorithme étend l'algorithme de Monte Carlo échantillonné à partir d'un espace de dimension variable par intégration de certains mouvements de saut réversible proposé par Green (1995). Pour des précisions sur l'algorithme, voir la documentation et le logiciel sur lesquels s'appuie l'article. Ils sont consultables à <http://www.stat.tamu.edu/~liang>.

2.2 Estimation pour population finie au moyen de

réseaux neuronaux bayésiens

2.2.1 Calcul de la moyenne des modèles bayésiens

À la présente sous-section, nous examinons certains résultats élémentaires du calcul de la moyenne des modèles bayésiens et montrons un théorème pour les modèles BNN qui représentent le fondement théorique des estimateurs décrits à la section 2.2. Supposons que nous voulions estimer la quantité $p(\theta^V)$, qui est une fonction de Δ et de θ . L'estimateur bayésien de $p(\theta^V)$ peut s'écrire

$$E_{\pi} p(\theta^V) = \sum_{K=1}^K p(V_K | D) \int p(\theta^K, V_K | D) \pi(\theta^K | V_K, D) d\theta^K. \quad (5)$$

où K représente le nombre total de modèles envisagés, θ^K représente les paramètres associés au modèle V_K et $\pi(\theta^K | V_K, D)$ représente la densité de probabilité a posteriori de θ^K sachant le modèle V_K . Madigan et Raftery (1994) soutiennent que, pour cet estimateur, la moyenne des modèles bayésiens (calcul de la moyenne sur l'ensemble des modèles de cette façon) tient compte de l'incertitude du modèle et qu'elle fournit une meilleure capacité de prédiction, mesurée par la règle de notation logarithmique, que tout modèle unique V_K . Voir Hoeting, Madigan, Raftery et Volinsky (1999) pour un tutoriel sur le calcul de la moyenne des modèles bayésiens.

γ_{j0} représente le terme de biais de la j° unité cachée, $\gamma_{j1}, \dots, \gamma_{jp}$ représentent les poids appliqués aux connexions allant des unités d'entrée à la j° unité cachée; et $\psi(\cdot)$ représente la fonction d'activation. Les fonctions sigmoïde et tangente hyperbolique sont deux choix fréquents pour la fonction d'activation. Nous fixons $\psi(z) = \tanh(z)$ pour tous les exemples présentés dans l'article.

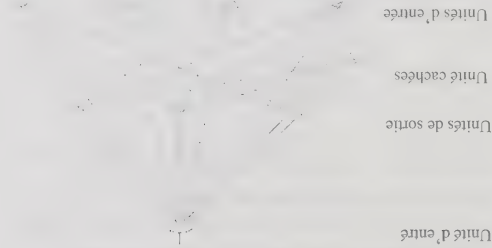


Figure 1. Un réseau neuronal feed-forward à une couche cachée entièrement connecté, avec quatre unités d'entrée, trois unités cachées et une unité de sortie. Les flèches indiquent la direction de la progression des données.

Soit Δ le vecteur formé de tous les indicateurs du modèle (3). Notons que Δ spécifie la structure du réseau correspondant. Soit $\mathbf{a} = (a_0, a_1, \dots, a_p)$, $\mathbf{\beta} = (\beta_1, \dots, \beta_M)$, $\gamma_j = (\gamma_{j0}, \dots, \gamma_{jp})$, $\gamma = (\gamma_1, \dots, \gamma_M)$, et $\boldsymbol{\theta} = (\boldsymbol{a}_V, \boldsymbol{\beta}_V, \gamma_V, \sigma_V^2)$, où $\mathbf{a}_V, \boldsymbol{\beta}_V$ et γ_V représentent les sous-ensembles non nuls de $\mathbf{a}, \mathbf{\beta}$ et γ , respectivement. Donc, le modèle (3) est spécifié complètement par le tuple $(\boldsymbol{\theta}, \Delta)$. Par souci de simplicité, dans la suite, nous utilisons $\boldsymbol{\theta}_V$ pour représenter un modèle BNN et $\hat{g}(x, \boldsymbol{\theta}_V)$ pour redénoter la fonction $\hat{g}(x', \mathbf{a}, \boldsymbol{\beta}, \gamma)$. En outre, nous posons que $\boldsymbol{\theta}_V = (\boldsymbol{\theta}, \Delta)$ et utilisons $\boldsymbol{\theta}_V$ et $(\boldsymbol{\theta}, \Delta)$ de façon interchangeable. Pour réaliser une analyse bayésienne pour le modèle (3), nous avons les lois a priori suivantes : $a_i \sim N(0, \sigma_a^2)$ pour $a_i \in \mathbf{a}_V$; $\beta_j \sim N(0, \sigma_\beta^2)$ pour $\beta_j \in \boldsymbol{\beta}_V$; $\gamma_{ji} \sim N(0, \sigma_\gamma^2)$ pour $\gamma_{ji} \in \gamma_V$; et $f(\sigma_i^2) \sim 1/\sigma_i^2$. Le nombre total de connexions effectives dans Δ est $m = \sum_{j=0}^p I_{\Delta_j}^{\Delta_j} + \sum_{j=1}^M I_{\Delta_j}^{\Delta_j} I_{\Delta_j}^{\Delta_j}$, où $\delta(\sum_{j=0}^p I_{\Delta_j}^{\Delta_j}) + \sum_{j=1}^M I_{\Delta_j}^{\Delta_j} I_{\Delta_j}^{\Delta_j}$ et 0 autrement. Le modèle Δ est assujéti à une probabilité a priori qui est proportionnelle à la masse placée sur m par une loi de Poisson tronquée (λ) avec taux λ .

$$P(\Delta) = \begin{cases} \frac{1}{Z} \frac{\lambda^m}{m!}, & m = 3, 4, \dots, U \\ 0, & \text{autrement} \end{cases}$$

où $U = (M + 1)(p + 1) + M$ est le nombre de connexions dans le modèle complet dans lequel tous les $I_{\Delta_j}^{\Delta_j} = 1$; et $Z = \sum_{\Delta \in \mathcal{A}} \lambda^m / m!$. Ici, nous représentons par \mathcal{A} l'ensemble

Dans notre modèle BNN, la fonction $g(\cdot)$ du modèle (2) est approximée par une fonction de la forme

$$\hat{g}(x', \mathbf{a}, \boldsymbol{\beta}, \gamma) = a_0 I_{a_0} + \sum_{i=1}^p x'_i a_i I_{a_i} + \sum_{j=1}^M \beta_j I_{\beta_j} \psi \left(\sum_{i=1}^p x'_i \gamma_{ji} I_{\gamma_{ji}} + \gamma_{j0} I_{\gamma_{j0}} \right), \quad (3)$$

où $I_{\Delta_j}^{\Delta_j}$ est une fonction indicateur qui indique l'efficacité de la connexion Δ_j ; M représente le nombre maximal d'unités cachées qui est spécifié par l'utilisateur; a_0 représente le terme de biais de l'unité de sortie; a_1, \dots, a_p représentent les poids appliqués aux connexions d'entrée à l'unité de sortie; β_1, \dots, β_M représentent les poids appliqués aux connexions d'entrée à l'unité de sortie; $\gamma_{11}, \dots, \gamma_{jp}$ représentent les poids appliqués aux connexions allant des unités cachées à l'unité de sortie; et γ_{j0} représente le terme de biais de l'unité de sortie.

progression de l'information vers l'avant). La figure 1 illustre un réseau neuronal feed-forward avec une couche cachée. Il comprend quatre types d'unités, à savoir des unités de biais, des unités d'entrée, des unités cachées et des unités de sortie. L'unité à laquelle sont présentées les caractéristiques d'entrée est appelée unité d'entrée. L'unité de biais est un type particulier d'unité d'entrée ayant une entrée constante; disons, 1. L'unité où se forme la sortie du réseau est appelée unité de sortie. L'unité cachée est appelée ainsi parce que son entrée et sa sortie sont utilisées uniquement pour des connexions internes et ne sont pas accessibles au monde extérieur. Dans un réseau neuronal feed-forward, chaque unité cachée traite indépendamment les valeurs qui lui sont fournies par les unités de la couche précédente, puis présente sa sortie aux unités de la couche suivante en vue d'un traitement supplémentaire. Plusieurs auteurs (Cybenko 1989; Funahashi 1989; Hornik Stinchcombe et White 1989) ont montré que les réseaux neuronaux sont des approximations universels en ce sens qu'un réseau neuronal feed-forward à une couche cachée avec unités de sortie linéaires peut approximer arbitrairement bien toute fonction continue sur des ensembles compacts en augmentant le nombre d'unités cachées. En ce qui concerne la régression de données d'enquête, il s'agit d'un avantage important des modèles de réseau neuronal par rapport aux autres modèles de régression. Dans la littérature traitant de la régression des données d'enquête, qu'elle soit assistée par modèle ou fondée sur un modèle, une attention considérable est généralement accordée aux conséquences de l'erreur de spécification du modèle. Le modèle de réseau neuronal permet d'éviter ces considérations, en partie grâce à sa propriété spécifique d'approximation universelle. À la section 2.2.1, nous montrons que, comme la taille d'échantillon est grande, la fonction de régression inconnue $g(\cdot)$ dans (2) peut être bien approximée par les modèles BNN, laquelle que soit la forme réelle de la fonction $g(\cdot)$. Essentiellement, le BNN rentre dans la classe des méthodes guidées par les données.

modèles de travail pour établir les équations de calage. L'approche du calage fondé sur un modèle peut être qualifiée d'approche « assistée par modèle », parce que, si l'efficacité de l'estimateur calé d'après un modèle dépend de la validité du modèle, il n'en est pas ainsi de sa convergence.

La littérature sur les sondages témoigne d'une tendance croissante à utiliser la régression non linéaire et non paramétrique. Au lieu du modèle (1), on considère

$$y_i = g(x_i) + \epsilon_i,$$

où la fonction de régression $g(\cdot)$ peut correspondre à toute fonction lisse arbitraire. Dorfman (1992) estime g en utilisant l'estimateur à noyau \hat{g} de Nadaraya-Watson pour obtenir l'estimateur ou prédicteur fondé sur un modèle qui suit de la moyenne de population finie

$$\hat{y}_N = N^{-1} \left\{ \sum_{i=1}^n y_i + \sum_{N=n+1}^N \hat{g}(x_i) \right\},$$

où il est supposé sans perte de généralité que l'échantillon est formé des n premiers éléments de la population. Kuk (1993) utilise la méthode du noyau pour estimer la loi conditionnelle de y sachant x comme moyen d'intégrer l'information auxiliaire dans l'estimation de la loi de y en population finie. Pour le cas du scalaire x , Bredt et Opsomer (2000) estiment g par la régression polynomiale locale avec poids de sondage intégrés pour tenir compte du plan de sondage utilisé et proposent un estimateur par différence généralisée,

$$\hat{y}^{LP} = N^{-1} \left\{ \sum_{i=1}^n \frac{\pi_i}{y_i - \hat{g}(x_i)} + \sum_N \hat{g}(x_i) \right\} = N^{-1} \left\{ \sum_{i=1}^n w_i y_i \right\},$$

où π_i est la probabilité de sélection dans l'échantillon. On peut montrer que les poids w_i sont calés sur les totaux de x jusqu'au q^{e} ordre, où q est l'ordre du polynôme local. Par conséquent, \hat{y}^{LP} est exactement sans biais par rapport au modèle si la fonction de régression réelle est un polynôme de degré q ou inférieur. Bredt et Opsomer (2000) montrent aussi que \hat{y}^{LP} est asymptotiquement sans biais par rapport au plan de sondage et converge sous des conditions faibles. Pour d'autres discussions des méthodes non linéaires et non paramétriques, voir Valliant, Dorfman et Royall (2000) (chapitre 11).

Dans le présent article, nous appliquons une autre méthode par la régression non linéaire, celle du réseau neuronal bayésien (BNN pour *Bayesian Neural Network*), Le BNN a l'avantage important de permettre de traiter facilement la sélection de variables auxiliaires multivariées et celle du modèle, ce qui n'est pas le cas pour de nombreuses autres techniques non linéaires et non paramétriques. Les premiers BNN ont été introduits par Buntline et Weigend (1991) et par MacKay (1992), et leur

La suite de l'article est présentée comme suit. À la section 2, nous décrivons les modèles BNN et les estimation linéaire. À la suite de l'article est présentée comme suit. À la section 2, nous décrivons les modèles BNN et les estimation linéaire. À la section 3, nous présentons les résultats numériques pour un exemple de population finie avec deux jeux de variables auxiliaires et les comparaisons à divers modèles fondés sur la régression linéaire. À la section 4, nous présentons les résultats numériques pour un autre exemple de population finie et démontrons comment on peut appliquer une méthode de validation croisée pour déterminer les paramètres des modèles BNN. À la section 5, nous concluons par une brève discussion.

2. ESTIMATION POUR POPULATION FINIE AU MOYEN DE RÉSEAUX NEURONAUX BAYÉSIENS

2.1 Modèles de réseau neuronal bayésien (BNN)

Supposons que nous disposions de paires de données $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ générées à partir de la relation

$$y_i = g(x_i) + \epsilon_i,$$

où $y_i \in R^1$, $x_i = (x_{i1}, \dots, x_{ip}) \in R^p$, $g(\cdot)$ est la fonction de régression réelle de forme inconnue et $\epsilon_i/\sigma \sim t(\nu)$ avec $\nu > 2$ un nombre de degrés de liberté connu de la loi t . Ici, $g(\cdot)$ peut être fortement non linéaire et σ est un paramètre d'échelle inconnu. Nous utilisons la loi t de Student au lieu de la loi gaussienne habituelle pour modéliser les perturbations, afin de pouvoir tenir compte des observations extrêmes souvent présentes dans les données provenant d'enquêtes sociales par sondage.

Avant de décrire notre modèle BNN, commençons par décrire brièvement les réseaux neuronaux feed-forward (à

Une étude de l'estimation pour population finie au moyen de réseaux neuronaux bayésiens

FAMING LIANG et ANTHONY YUNG CHEUNG KUK¹

RÉSUMÉ

Nous étudions l'utilisation de réseaux neuronaux bayésiens pour l'estimation pour population finie. Nous proposons des estimateurs de la moyenne de population finie et de son erreur quadratique moyenne. Nous proposons aussi d'utiliser la loi t de Student pour modéliser les perturbations afin de pouvoir utiliser les observations extrêmes souvent présentes dans les données provenant d'enquêtes sociales par sondage. Les résultats numériques montrent que les réseaux neuronaux bayésiens améliorent significativement l'estimation pour population finie comparativement aux méthodes fondées sur la régression linéaire.

MOTS CLÉS : Calcul de la moyenne des modèles bayésiens; réseau neuronal bayésien, méthode de Monte Carlo évolutive; population finie; méthode de Monte Carlo

1. INTRODUCTION

Dans le contexte des enquêtes par sondage, il est très fréquent de recourir à l'estimation par la régression pour intégrer l'information auxiliaire sur la population (Cochran

$$y_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1)$$

où y_i est la variable d'intérêt pour le i^{e} élément d'une population, $\mathbf{x}_i' = (x_{i1}, \dots, x_{ip})$ est le vecteur de variables auxiliaires associé à y_i , $\beta_0, \beta_1, \dots, \beta_p$ sont les coefficients de régression et ε_i est la perturbation indépendante de moyenne nulle et de variance commune. Bien qu'il donne généralement de bons résultats, ce modèle présente plusieurs limites intrinsèques. Premièrement, il est spécifié linéairement et ne peut donc refléter certains types de relations non linéaires, propriété qui est essentielle dans certaines applications. Deuxièmement, l'estimation par les moindres carrés, qui est très répandue pour le modèle (1), pourrait ne pas être fiable en présence de colinéarité des variables auxiliaires. Le cas échéant, il est nécessaire de recourir à des techniques, comme la réduction du nombre de conditions (Bankier 1990) ou la régression ridge (Bardsley et Chambers 1984) et à diverses procédures de sélection des variables (Silva et Skinner 1997) pour améliorer la prédictivité médiocre du modèle. Troisièmement, la présence de valeurs aberrantes peut affecter gravement l'estimation par les moindres carrés.

Certaines tentatives ont été faites en vue de rendre les estimateurs moins dépendants du modèle linéaire (1). Firth et Bennett (1998) précisent une condition « de calage interne du biais » suffisante sous laquelle un estimateur

fondé sur un modèle est automatiquement convergent par rapport au plan de sondage, quelle que soit la qualité de l'ajustement du modèle sous-jacent à la population. La condition est satisfaite par certains estimateurs fondés sur des modèles linéaires, certains modèles linéaires généralisés à lien canonique, ainsi que les estimateurs par régression non paramétrique construits d'après ces modèles par une méthode particulière d'ajustement local de la vraisemblance. Le biais peut également être calé extérieurement s'il ne peut l'être intérieurement. Chambers, Dorfman et Wehrly (1993) parlent d'un prédicteur de la moyenne de population fondé sur un modèle linéaire hétéroscédastique et corrigent le biais moyen d'une régression non paramétrique. Kuk et Welsh (2001) proposent une approche fondée sur un modèle rendu robuste consistant à ajuster d'abord un modèle de travail en utilisant des méthodes robustes, puis à estimer non paramétriquement les lois conditionnelles des résidus sachant \mathbf{x} pour tenir compte de l'écart par rapport au modèle local ou des valeurs aberrantes dans des régions localisées.

Un autre moyen d'intégrer l'information auxiliaire dans les valeurs abstraites de données est d'utiliser des modèles linéaires généralisés en utilisant les valeurs ajustées sous ces

HINKLEY, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19, 285-292.

HOLT, D., et SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, A, 142, 33-46.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

KALTON, G. (1977). Practical methods for estimating survey sampling errors. *Bulletin of the International Statistical Institute*, 47, 495-514.

LAZZARONI, L.C., et LITTLE, R.J.A. (1998). Random effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

LIN, X., et ZHANG, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society*, B, 61, 381-400.

LITTLE, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

MILLER, R.G. (1974). An unbalanced jackknife. *Annals of Statistics*, 2, 880-891.

SÄRANDAL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

SHAO, J., et WU, C.F.J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models. *Annals of Statistics*, 15, 1563-1579.

SHAO, J., et WU, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

U.S. CENSUS (1990). Dept. of Commerce, Census of Population and Housing. [United States]: fichier de l'échantillon de microdonnées à grande diffusion à 5 %. 3^e éditions. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 1995. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributeur], 1996.

ZHENG, H., et LITTLE, R.J.A. (2003). Penalized spline model-based proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

ZHENG, H., et LITTLE, R.J.A. (2004). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. À paraître dans *Journal of Official Statistics*.

les dénombrements d'UPF sont connus pour l'échantillon, mais non pour la population dans son ensemble, les estimations fondées sur un modèle des dénombrements inconnus peuvent encore fournir des estimations valables de la moyenne de population si le modèle suit les dénombrements réels d'UPF de façon suffisamment précise. Nous avons traité le modèle reliant ces dénombrements à la variable auxiliaire de façon paramétrique, mais il pourrait être spécifique de façon non paramétrique sans trop de difficulté.

Nous sommes convaincus que les modèles non paramétriques mixtes à fonction *p*-spline peuvent être appliqués à des plans de sondage plus complexes, tels que les plans stratifiés et les plans à plusieurs degrés. Nous pensons aussi que notre méthode peut être généralisée sans trop d'efforts à des résultats binaires ou ordinaux.

REMERCIEMENTS

La présente étude a été financée par la bourse DMS 0106914 de la National Science Foundation.

BIBLIOGRAPHIE

BRUMBACH, B.A., RUPPERT, D. et WAND, M.P. (1999). Comment to variable selection and function estimation in additive nonparametric regression using data-based prior. *Journal of the American Statistical Association*, 94, 794-797.

COUL, B.A., SCHWARTZ, J. et WAND, M.P. (2001) Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics*, 2(3), 337-349.

ELLIOTT, M.R., et LITTLE, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

FIRTH, D., et BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, B, 60, 3-21.

GOSH, M., et MEEDEN, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.

Tableau 4 Estimation de la variance et des taux empiriques de couverture des IC à 95 % en utilisant une fonction p -spline et des dénombrements d'UPB estimés, population simulée avec des erreurs normales

	$\sigma = 0,1$ et $\tau = 0,2$				$\sigma = 0,2$ et $\tau = 0,1$			
	Variance	Taux de	Variance	Taux de	Variance	Taux de	Variance	Taux de
NULL	90	76	91,8	91,8	46	50	46	93,2
LINDOWN	93	74	90,4	90,4	43	43	46	95,6
EXP	85	72	93,0	94,8	55	55	56	96,2
SINE	110	98	94,8	94,8	50	50	55	97,6

Tableau 5 Estimation de la variance et des taux empiriques des IC à 95 % en utilisant une fonction p -spline et des dénombrements d'UPB estimés, population simulée avec des erreurs lognormales

	$\sigma = 0,1$ et $\tau = 0,2$				$\sigma = 0,2$ et $\tau = 0,1$			
	Variance	Taux de	Variance	Taux de	Variance	Taux de	Variance	Taux de
NULL	105	84	91,8	91,8	95	99	99	94,8
LINDOWN	103	98	94,4	94,4	110	102	102	94,4
EXP	81	79	94,6	94,6	87	83	83	94,2
SINE	110	150	96,4	96,4	91	130	130	95,8

Pour l'étude en simulation au moyen des données de l'EMGD à 5 %, la moyenne simple a un biais = -50,9 et une REQM = 2 600, et le modèle non paramétrique mixte à fonction p -spline (MMP) a un biais égal à -41,9 et une REQM = 2 153. Donc, les deux méthodes produisent un biais faible et l'estimateur fondé sur le modèle a une REQM inférieure de 17 % à celle de l'estimateur de la moyenne simple. Cet accroissement de l'efficacité est dû au fait que le revenu moyen des ménages diminue à mesure que le nombre de familles augmente dans les UPB (Figure 1). La méthode MMP exploite cette relation dans ses prédictions.

6. DISCUSSION

Les méthodes d'inférence fondées sur un modèle paramétrique utilisées antérieurement ont été critiquées principalement parce qu'un biais par rapport au plan de sondage important peut être introduit si le modèle est spécifié incorrectement. Dans nos modèles non paramétriques, l'hypothèse de linéarité est remplacée par une hypothèse beaucoup plus faible de relation à variation lisse. Par conséquent, les estimateurs fondés sur un modèle sont plus robustes, leurs biais étant faibles pour diverses formes de population.

Les méthodes d'inférence fondées sur un modèle paramétrique utilisées antérieurement ont été critiquées principalement parce qu'un biais par rapport au plan de sondage important peut être introduit si le modèle est spécifié incorrectement. Dans nos modèles non paramétriques, l'hypothèse de linéarité est remplacée par une hypothèse beaucoup plus faible de relation à variation lisse. Par conséquent, les estimateurs fondés sur un modèle sont plus robustes, leurs biais étant faibles pour diverses formes de population.

La méthode bayésienne empirique, ainsi que les méthodes du jackknife et BRR donnent toutes une bonne couverture des intervalles de confiance, lesquels sont plus étroits que ceux produits par les méthodes traditionnelles. Cependant, nous nous attendons à ce que la méthode bayésienne empirique soit sensible aux hypothèses du modèle concernant les composantes de la variance (par exemple, variances à l'intérieur des UPB constantes). Si

Tableau 2
Biais et REQM empiriques pour les modèles MMP, HT, GR et MMP avec N_h estimé sous plans d'échantillonnage avec probabilités inégales

	(×10 ⁻³)				MMP				Horvitz-Thompson				Linéaire assisté				MMP avec N _h																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																						
	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales	LINDOWN	EXP	SINE	Erreurs	normales

par un biais empirique par rapport au plan de sondage très faible. Nous pensons qu'il en est ainsi parce que les fonctions de moyenne souples donnent de bonnes prédictions des moyennes d'UPE.

Au tableau 3, nous comparons les estimations ponctuelles et la couverture des intervalles de confiance à 95 % calculée par trois méthodes d'estimation de la variance pour l'estimateur MMP, à savoir la méthode fondée sur un modèle bayésien empirique, la méthode du jackknife et la méthode BRR. La méthode bayésienne empirique est généralement satisfaisante, mais a tendance à sous-estimer la variance réelle de l'estimateur MMP, ce qui donne lieu à un sous-couverture dans certains cas. Les méthodes jackknife et BRR ont tendance à produire des estimations plus robustes de la variance. En général, l'estimateur MMP donne des estimations plus efficaces que les estimateurs HT et linéaires assistés par modèle habituels et des estimations fondées sur le plan de sondage satisfaisantes.

ainsi qu'à l'estimateur linéaire assisté par modèle. Pour obtenir cette plus grande efficacité, il faut connaître l'information sur le plan de sondage, y compris les probabilités $\pi_{i,h}$ et les dénombrements d'UPE N_h pour l'ensemble de la population. Si l'on utilise des dénombrements d'UPE estimés N_h à la place de N_h , l'estimateur résultant est moins efficace qu'avec le N_h connu, mais l'estimateur MMP peut encore surpasser l'estimateur HT quand la fonction moyenne n'est pas constante. Les comparaisons sur les populations dont les erreurs à l'intérieur des UPE sont normales ou lognormales produisent des résultats comparables.

Le tableau 2 montre une amélioration comparable pour la méthode MMP dans le cas de l'échantillonnage avec probabilités inégales, ce qui donne à penser que l'élément essentiel de l'accroissement de l'efficacité est la meilleure prédiction donnée par les modèles non paramétriques. Les tableaux 1 et 2 font penser l'un et l'autre que les estimateurs fondés sur un modèle à fonction p -spline sont caractérisés

Tableau 1
Biases et REQM empiriques des estimateurs MMP, HT, GR et MMP avec N_h estimé sous plans d'échantillonnage avec probabilités égales

		(×10 ⁻³)						Linéaire assisté par					
		MMP			Horvitz-Thompson			modèle			MMP avec N_h estimé		
		BIAS	REQM	BIAS	REQM	BIAS	REQM	BIAS	REQM	BIAS	REQM	BIAS	REQM
Erreurs	normales	1,1	29,7	0,8	30,0	0,8	29,9	29,9	1,3	30,1	30,4	2,3	30,4
	$\tau = 0,2$	−4,4	29,1	−9,4	53,0	−9,5	36,7	−4,3	29,1	34,3	22,3	5,5	20,6
	$\sigma = 0,1$	4,8	32,5	2,1	42,0	−0,3	35,9	5,2	34,3	26,4	23,4	0,4	20,5
Erreurs	normales	5,7	22,0	6,6	22,5	6,6	22,1	22,1	5,5	22,3	20,6	1,6	20,6
	$\tau = 0,1$	0,9	23,1	1,9	50,3	−4,2	31,7	0,4	23,4	26,4	20,6	8,0	26,4
	$\sigma = 0,2$	7,0	22,3	6,5	34,9	3,8	26,4	8,0	26,4	26,4	23,4	0,4	20,5
Erreurs	lognormales	1,7	32,3	0,9	32,3	0,7	32,3	32,3	1,5	32,5	32,0	3,2	32,0
	$\tau = 0,2$	−0,6	28,4	−5,9	51,5	−6,9	36,4	−0,3	28,5	35,0	30,8	6,4	33,1
	$\sigma = 0,1$	6,9	33,8	1,5	43,7	−1,9	39,0	−3,1	35,0	35,0	30,8	9,1	30,8
Erreurs	lognormales	8,5	30,5	9,6	31,3	9,2	31,0	31,0	9,1	30,8	30,8	9,1	30,8
	$\tau = 0,1$	3,9	29,0	6,8	53,8	1,0	34,4	34,4	3,7	29,4	29,4	3,7	29,4
	$\sigma = 0,2$	−2,9	30,1	−8,9	44,7	−12,0	38,4	−3,8	35,9	35,9	30,8	9,1	30,8

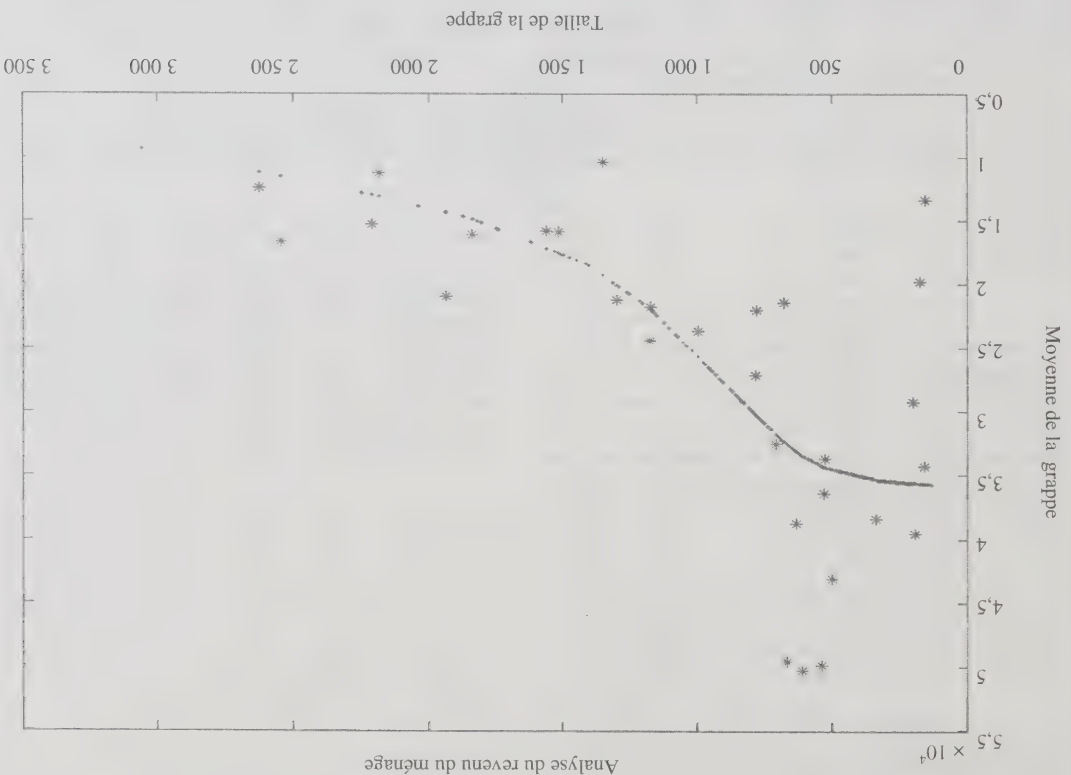


Figure 1. Courbe de régression p -spline (trait pointillé) et revenu moyen du ménage (étoiles) dans les UPE échantillonnées

Nous tirons cinq échantillons à deux degrés, chacun comprenant 30 UPE et 20 USE (familles) tirés de chaque UPE sélectionnée. L'échantillonnage de premier degré est un échantillonnage PPT systématique où les mesures de taille sont égales aux dénombrements (nombre d'unités secondaires) des UPE. L'échantillon de deuxième degré est un échantillonnage aléatoire simple avec probabilités de sélection proportionnelles à l'inverse des probabilités de sélection de premier degré. Pour estimer la moyenne, nous utilisons des dénombrements réels d'UPE comme variable x_h , dont les valeurs sont proportionnelles aux probabilités de sélection de premier degré. Nous appliquons le modèle non paramétrique mixte à fonction p -spline formulé en (2). Nous utilisons 10 centiles d'échantillon également espacés des dénombrements d'UPE comme nœuds de la p -spline.

Le tableau 1 semble indiquer que les méthodes MIMP donnent des estimateurs à biais faible. Dans le cas de l'échantillonnage avec probabilités égales, l'estimateur MIMP est presque aussi efficace que l'estimateur HT quand la fonction de moyenne f est constante. Dans les cas plus généraux, tels que NULL et LINDOWN, où f est linéaire, mais non constante, la méthode linéaire assistée par modèle et la méthode MIMP sont comparables et toutes deux plus efficaces que l'estimateur HT en ce qui concerne la racine de l'erreur quadratique moyenne. Pour les populations EXP et SINE, dont les fonctions de moyenne ne sont pas linéaires, la méthode MIMP est supérieure à l'estimateur HT

L'erreur quadratique moyenne (REQM) pour quatre

5.2 Résultats

où

$$E(\tilde{T} | \hat{N}_h, \pi_h, M_h) = \sum_{h=m+1}^m (N_h - n_h) \pi_h + \sum_{H=m+1}^H \hat{N}_H \pi_H$$

et

$$\text{Var}(\tilde{T} | \hat{N}_h, \pi_h, M_h) \approx \sigma^2 (\tilde{N}^p X^p (X^T X + \Delta)^{-1} X^T \tilde{N}^p),$$

$\tilde{N}^p = [(N_1 - n_1) \dots (N_m - n_m) \dots \hat{N}_{m+1}]$ et X^p, X^m et Δ sont définis de la même façon que dans (3).

Si les modèles de μ_h et N_h sont tous deux correctement spécifiés, nous pouvons estimer la variance susmentionnée conformément aux modèles correspondants.

5. SIMULATIONS

5.1 Plan de simulation

Nous avons réalisé deux simulations pour comparer la méthode de pondération par la probabilité inverse, la méthode assistée par modèle (Sæmndal et coll. 1992) et la méthode MIMP dans le cas de l'échantillonnage à deux degrés.

Dans notre première simulation, nous générons des populations artificielles dont les fonctions de moyenne $f(\pi_{1,h})$ des probabilités de sélection au premier degré sont différentes. Nous simulons quatre fonctions de moyenne différentes, à savoir 1) NUTL, une fonction constante; 2) LINDOWN, une fonction linéairement décroissante; 3) EXP, une fonction exponentiellement croissante et 4) SINE, une fonction sinusoïdale.

Nous simulons deux combinaisons de valeurs des composantes de la variance: 1) $\sigma = 0,1$ et $\tau = 0,2$; 2) $\sigma = 0,2$ et $\tau = 0,1$. Enfin, nous simulons uniquement des erreurs normales autour des fonctions de moyenne, ainsi que des erreurs normales et lognormales à l'intérieur des UPE.

La population comprend 500 UPE et, au premier degré, 48 UPE sont échantillonnées systématiquement avec probabilité proportionnelle à la taille (PPT) à partir d'une liste aléatoirement ordonnée. Les tailles des UPE sont distribuées uniformément, les valeurs variant de 4 à environ 400. Le nombre d'USE dans chaque UPE est généré à partir d'une loi de moyenne égale à 1,05 fois la mesure de taille et d'erreurs lognormales avec un écart-type de 30.

Nous étudions deux types de plans d'échantillonnage à deux degrés: 1) l'échantillonnage aléatoire simple (cas) à l'intérieur des UPE avec probabilités de sélection proportionnelles à l'inverse des probabilités de sélection au premier degré, aboutissant à une probabilité de sélection égale pour toutes les USE, 2) un échantillonnage aléatoire simple à l'intérieur des UPE avec le même taux d'échantillonnage pour toutes les UPE échantillonnées, si bien que les probabilités de sélection résultantes pour les USE comprises dans l'UPE h soient proportionnelles à $\pi_{1,h}$.

Pour chaque échantillon tiré sous les deux plans d'échantillonnage, nous appliquons les méthodes qui suivent.

A. L'estimateur HT.

B. La méthode d'estimation assistée par modèle. Nous utilisons un modèle linéaire de régression de la variable de résultat y_{hi} sur les probabilités de sélection au premier degré qui sont traitées comme des informations de niveau élémentaire. Nous calculons l'estimateur GR au moyen de la formule donnée à la section 1.

C. La méthode MIMP, avec les probabilités de sélection au premier degré $\pi_{1,h}$ comme covariable. Nous utilisons 20 centiles éaux de $\pi_{1,h}$ des UPE échantillonnées comme nœuds pour la p -spline de régression.

D. La méthode MIMP avec les moyennes d'UPE μ_h estimées de la même façon qu'en C, mais en utilisant des nombres estimés d'UPE d'après une régression linéaire simple de N_h sur les mesures de taille, qui sont proportionnelles à $\pi_{1,h}$. Nous réalisons cette partie de la simulation pour étudier la méthode décrite à la section 4.

Les estimations de \bar{Y} obtenues par les méthodes A à D sont calculées pour chacun des 500 échantillons tirés par répétition à partir des populations artificielles (chaque population artificielle n'est générée qu'une seule fois). Pour l'estimateur MIMP, nous calculons les estimateurs de la variance par la méthode bayésienne empirique, la méthode du jackknife ($K = 8$) et la méthode BRR pour chaque échantillon répété. Les estimations moyennes de la variance de l'estimateur MIMP et du taux de couverture correspondant à un intervalle de confiance à 95 % sont utilisées pour juger de la qualité de l'inférence. Pour la méthode D, nous étudions le biais empirique de l'estimateur de la variance fondé sur un modèle décrit à la section 4, ainsi que les taux de couverture des intervalles de confiance connexes.

Dans la seconde étude en simulation, nous tirons des échantillons de données sur le revenu du ménage provenant de l'échantillon de microdonnées à grande diffusion (EMGCD) à 5 % pour l'Etat du Michigan au recensement des Etats-Unis de 1990, que nous traitons comme une population finie. Cette simulation est plus réaliste que la précédente en ce sens que les valeurs de la variable de résultat proviennent d'une distribution réelle plutôt que simulée. Les UPE que nous simulons sont fondées sur les grappes géographiques naturelles appelées « secteurs de microdonnées à grande diffusion » (SMGD), qui sont habituellement des comtés et des localités. Il existe 67 SMGD dans l'EMGCD à 5 % du Michigan pour lesquels les nombres de familles varient d'environ 1 300 à plus de 10 000. Nous augmentons le nombre d'UPE disponibles en divisant chaque SMGD par cinq, ce qui donne 335 UPE. Le nombre d'USE varie de 134 à 3 058. La figure 1 donne le nuage de points d'un échantillon du revenu moyen du ménage en fonction des tailles d'UPE, ainsi que la courbe de régression $f(x)$.

La variance a posteriori bayésienne empirique de

$$\hat{Y} = \hat{I}_{\text{précise}} / N \text{ est}$$

$$\text{Var}(\hat{Y} | \sigma^2, \sigma_b^2, \tau^2, X, X^p) =$$

$$\sigma^2 (N^p X^p (X^T X + \Delta)^{-1} X^p N^p) / N^2.$$

3.2 Méthode du jackknife

Nous élaborons un estimateur de la variance par le jackknife pour l'estimateur MME. Nous produisons les répliques jackknife en divisant l'ensemble d'UPF en G sous-groupes égaux et en calculant la g^e pseudovaleur selon $\hat{Y}_g = G\bar{Y} - (G-1)\bar{Y}^{(g)}$, où \bar{Y} est l'estimateur MME original et $\bar{Y}^{(g)}$ est le même estimateur calculé pour l'échantillon réduit obtenu en excluant les éléments provenant des UPF comprises dans le g^e sous-groupe.

L'estimation de la variance de \bar{Y} par le jackknife est

$$v(\bar{Y}) = \frac{1}{G(G-1)} \sum_{g=1}^G (\bar{Y}_g - \bar{Y})^2,$$

où $\bar{Y}_g = \sum_{i \in g} \hat{Y}_i / G$. Afin d'équilibrer la distribution des unités échantillonnées en m/G strates chacune de taille G avec des probabilités d'inclusion de premier degré semi-

biaises et nous construisons les G sous-groupes en sélectionnant aléatoirement un élément dans chaque strate. Pour réduire les calculs, nous ne recalculons pas les estimations $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ et $\hat{\tau}^2$ pour chaque réplique. Autrement dit, nous calculons des pseudovaleurs de $(\beta_0, \beta_1, \dots, \beta_{K+1}, u_1, \dots, u_m)^T$ d'après les composantes de la variance estimées à partir de l'échantillon complet.

Miller (1974), ainsi que Shao et Wu (1987, 1989) ont prouvé les propriétés asymptotiques de l'estimateur par le jackknife et de l'estimation de la variance par le jackknife dans le cas de la régression linéaire multiple. Zheng et Little (2004) ont fourni une justification théorique de la méthode du jackknife pour l'estimateur fondé sur un modèle à fonction p -spline dans le cas des plans de sondage à un degré. Les simulations numériques de la section 4 font penser que la méthode du jackknife susmentionnée donne aussi de bons résultats pour les plans d'échantillonnage à deux degrés. Une meilleure performance pourrait être obtenue en utilisant le jackknife pondéré proposé par Hinkley (1977).

3.3 Méthode des répliques répétées équilibrées

La méthode des répliques répétées équilibrées (BRR) peut être appliquée à des plans d'échantillonnage stratifiés avec tirage de deux unités dans chaque strate. Dans le cas des plans d'échantillonnage à une UPF par strate, il est fréquent de regrouper les strates (Kallott 1977) pour l'estimation de la variance par la méthode BRR. Dans notre

4. CAS OÙ CERTAINS DÉNOMBREMENTS D'UPF SONT INCONNUS

Aux sections 2 et 3, nous avons supposé que les nombres d'USE N_h étaient connus pour les UPF échantillonnées et non échantillonnées. À la présente section, nous examinons la situation où N_h n'est connu exactement que pour les UPF échantillonnées (annotées de 1 à m). Nous supposons aussi que les valeurs $M_h, h=1, \dots, H$ d'une variable auxiliaire prédictive de N_h sont connues pour l'ensemble de la population. Par exemple, les M_h pourraient être les nombres d'USE estimés d'après des sources externes, comme un recensement. Nous exécutons une régression de N_h sur M_h en utilisant les UPF échantillonnées et en remplaçant les dénominateurs N_h dans (3) pour les UPF non échantillonnées par les prédictions $\hat{N}_h, h=m+1, \dots, H$ faites d'après la régression. L'estimation résultante du total est

$$\tilde{T} = T_1 + \sum_{h=1}^m (N_h - n_h) \hat{p}_h + \sum_{h=m+1}^H \hat{N}_h \hat{p}_h.$$

L'estimation de la variance de \tilde{T} doit être intégrée à la variable supplémentaire de \hat{N}_h . Plus précisément, une expression de la variance fondée sur un modèle pour \tilde{T} est

$$\text{Var}(\tilde{T} | \pi_h, M_h) = \text{Var}(E(\tilde{T} | \hat{N}_h, \pi_h, M_h)) + E(\text{Var}(\tilde{T} | \hat{N}_h, \pi_h, M_h)),$$

Le plan d'échantillonnage est traité comme s'il était stratifié avec deux UPF contenant un certain biais, parce que le plan d'échantillonnage est traité comme s'il était stratifié avec deux UPF

$$\begin{bmatrix} 0 & \dots & \dots & 0 & {}^{+}(\chi\chi - H\chi) & \dots & {}^{+}(\chi - H\chi) & H\chi & 1 \end{bmatrix}$$

Nous pouvons interpréter le modèle (2) comme étant un modèle bayésien dans lequel les paramètres $n =$

modèle non paramétrique mixte à fonction p -spline qui suit pour l'inférence au sujet de la moyenne de population :

Modèle non paramétrique mixte à fonction p -spline (MMP):
$$\phi^u(x) = u^1, \dots, u^H, D^1, \dots, D^H, I^1, \dots, I^H$$
$$d^{+}(1-x)^{d+l}g\overline{Z}^{[=l]}+x^fg\overline{Z}^{[=f]}+0g=(g(x)f$$

ou $k_1 > k > k_2, \dots, p^{k_1}, \dots, p^{k_r}$ sont des entiers fixes, p^{k_1}, \dots, p^{k_r} sont les

traiter comme s'ils étaient fixes et à les estimer en même temps que les composantes de la variance σ^2 et τ^2 en

peut produire des estimations de f présentant trop de rugosité et de variabilité. Pour éviter de surajuster le modèle,

on peut pénaliser la rugosité de l'estimation \hat{f} en ajoutant un terme de pénalité à la somme des carrés des écarts, de sorte que la solution \hat{f}_{pen} minimise

$$\cdot \sum_K^{d+1} g(x) + \sum_m^{(q)} (g(x) - f(x))$$

Nous réalisons ceci dans le contexte du modèle en attribuant

une loi a priori normale $N^m(0, \sigma_z^2)$, et en posant que

Si $p = 1$, \hat{f} est linéaire par morceau et nous estimons les coefficients $\beta_0, \beta_{K+1}, \sigma^2, \sigma^2_{\beta}$ et τ^2 en ajustant le

modèle linéaire mixte :

$$y = X_1\beta + X_2u + e, \quad (2)$$

FONCTION p -SPLINE

$$= n^{-1} \cdot ({}^1\mathfrak{d} \cdot {}^0\mathfrak{d}) = \mathfrak{d}^{-1} \cdot ({}^{wuu}\kappa \cdot \dots \cdot \frac{{}^w n \cdot \dots \cdot {}^1 n \cdot {}^{1+X} \mathfrak{d} \cdot \dots \cdot {}^z \mathfrak{d}}{{}^{\ell} \ell! \kappa \cdot {}^{\ell!} \kappa}) = \kappa \quad \text{no}$$

où $t_h^* = \sum_{i=1}^m y_{hi}^* / \pi_{1,h}$, $h = 1, \dots, H$. En pratique, qu'elles soient

$$\hat{b} = \left(\sum_{h=1}^H x_h^T x_h^T / (\sigma_h^2 \pi_{1,h}) \right)^{-1} \sum_{h=1}^H x_h^T t_h^* / (\sigma_h^2 \pi_{1,h}),$$

(Särndal et coll. 1992). Nous estimons $\hat{\beta}$ par la régression pondérée par les probabilités

$$E(t_h | x_h) = x_h^T \beta, \text{Var}(t_h) = \sigma_h^2, h = 1, \dots, H$$

linéaire :

totaux d'UPE $t_h = \sum_{i=1}^{N_h} y_{hi}$ sont reliés à x_h selon un modèle niveau de l'UPE pour l'UPE h . Nous supposons que les auxiliaire disponible au niveau de l'UPE ou de l'USE. Dans pour les échantillons à deux degrés d'information HT Särndal, Swensson et Wretman (1992) ont discuté d'estimateurs assistés par modèle pour remplacer l'estimateur HT auxiliaires connues pour toutes les UPE de la population.

moyen d'un modèle de régression fondé sur des variables pour les UPE non échantillonnées peut être estimé à l'unique pour les UPE échantillonnées, mais où le N_h examinons la situation courante où N_h est connu UPE comprises dans la population. À la section 4, nous nombres N_h d'USE sont également connus pour toutes les $h = 1, \dots, H$. Aux sections 2 et 3, nous supposons que les sélection $\pi_{1,h}$ sont connues pour toutes les UPE Nous supposons dans tout l'article que les probabilités de

moyenne simple d'échantillon $\bar{y} = \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi} / \sum_{h=1}^H N_h$. pondérée par l'inverse des probabilités est égale à la égales pour toutes les USE. Dans ce cas, la moyenne bien que les probabilités globales de sélection π_h sont l'inverse des probabilités de sélection de premier degré, si sélection au deuxième degré sont proportionnelles à estimation de la taille de l'UPE et les probabilités de sélection au premier degré est proportionnelle à une plan d'échantillonnage adopté couramment, la probabilité d'unités (USE) dans l'ensemble de la population. Dans un l'unité dans l'UPE h et N est le nombre total connu $\sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi} / (\pi_{1,h} \pi_{2,h}) / N$, où y_{hi} est la valeur de Y pour de la moyenne d'un résultat Y est $\bar{y} =$ l'unité dans l'UPE h est $\pi_h = \pi_{1,h} \pi_{2,h}$, et l'estimateur HT sélection $\pi_{2,h}$. La probabilité globale de sélection de UPE échantillonnées étiquetées h avec la probabilité de secondaires d'échantillonnage (USE) est tiré à partir de un échantillon aléatoire simple (cas) de n_h parmi N_h unités incluses dans l'échantillon de 1 à m . À la deuxième étape, probabilités inégales $\pi_{1,h}$, $h = 1, \dots, H$. Numérotions les UPE (UPE) est tiré d'une population contenant H UPE, avec sous-ensemble de m unités primaires d'échantillonnage tilonnage à deux degrés. À la première étape, un Dans le présent article, nous considérons l'échan-

$$\hat{E}(\bar{Y} | y, x_h) = \frac{1}{N} \left(\sum_{h=1}^H n_h \bar{y}_h + (N_h - n_h) \bar{y}_h \right) + \sum_{h=m+1}^H N_h \bar{y}_h,$$

un modèle de \bar{Y} est donné par

où $\mu = (\mu_1, \dots, \mu_H)^T$, $\phi = (\phi_1, \dots, \phi_H)^T$ où μ_h est le résultat moyen dans l'UPE h , $\phi_h = x_h^T \beta$ et D est la matrice de covariance des moyennes d'UPE. L'estimateur fondé sur

$$(1) \quad \mu \sim N_H(\phi, D) \\ y_{hi}^* | \mu_h \sim N(\mu_h, \sigma^2)$$

de ce genre est :

Ces deux méthodes ne tiennent compte ni l'une ni l'autre des corrélations des résultats à l'intérieur des UPE. Ces moyennes d'UPE comme des effets aléatoires dans un modèle hiérarchique. Dans le cas où l'information au niveau de l'UPE x_h est disponible pour toutes les UPE, un modèle

$$\hat{y}_{hi}^* = x_{hi}^T \hat{\beta} = x_{hi}^T \left(\sum_{h=1}^H \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h} - \bar{y}_{hi} \right) / \pi_{hi},$$

l'échantillon. L'estimateur GR pour le total général est

$$\hat{b} = \left(\sum_{h=1}^H \sum_{i=1}^{N_h} x_{hi}^T x_{hi}^T / (\sigma_h^2 \pi_{hi}) \right)^{-1} \sum_{h=1}^H \sum_{i=1}^{N_h} x_{hi}^T y_{hi} / (\sigma_h^2 \pi_{hi}),$$

probabilités est

L'estimation de $\hat{\beta}$ par la régression pondérée par les

$$E(y_{hi} | x_{hi}) = x_{hi}^T \beta, \text{Var}(y_{hi}) = \sigma_h^2, h = 1, \dots, H, i = 1, \dots, N_h.$$

mation auxiliaire est modélisée par

$H, i = 1, \dots, N_h$. La relation entre le résultat et l'information auxiliaires pour l'USE i dans l'UPE h , $h = 1, \dots$ connue au niveau des USE, soit x_{hi} l'ensemble de variables

Dans le second cas, où l'information auxiliaire est l'estimateur convergent par rapport au plan de sondage.

et l'estimation de la moyenne est \bar{Y}^A / N . Le terme $\sum_{h=1}^H (t_h^* - \hat{t}_h) / \pi_{1,h}$ est un terme de calage du biais qui rend

$$\hat{Y}^A = \sum_{h=1}^H \hat{t}_h + \sum_{h=1}^H \frac{\pi_{1,h}}{(\hat{t}_h^* - \hat{t}_h)},$$

du total général est

qui précède. L'estimateur par la régression généralisée (GR) obtenues simplement par hypothèse (par exemple, σ_h proportionnel à une mesure de taille de la strate h) ou estimées, les estimations $\hat{\sigma}_h^2$ remplacent σ_h^2 dans la formule

Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés

HUI ZHENG et RODERICK J.A. LITTLE

RÉSUMÉ

Les échantillonneurs se prêtent souvent des approches d'inférence fondées sur un modèle, parce qu'ils craignent que soient commises des erreurs de spécification lorsque les modèles sont appliqués à de grands échantillons provenant de populations complexes. Nous soutenons que le paradigme de l'inférence fondée sur un modèle peut donner de très bons résultats dans les conditions d'enquête, à condition que les modèles soient choisis de façon à tenir compte du plan d'échantillonnage et d'éviter de faire des hypothèses fortes au sujet des paramètres. L'estimateur d'Horvitz-Thompson (HT) est un estimateur simple sans biais par rapport au plan de sondage du total de population finie. Du point de vue de la modélisation, l'estimateur HT fonctionne bien lorsque les ratios des valeurs de la variable de résultat et des probabilités d'inclusion sont interchangeables. Si cette hypothèse n'est pas satisfaisante, l'estimateur HT peut être très inefficace. Dans Zheng et Little (2003, 2004), nous avons utilisé des fonctions splines pénalisées (p -splines) pour modéliser des relations à variation lisse entre le résultat et les probabilités d'inclusion sous échantillonnage à un degré avec probabilité proportionnelle à la taille (PPT). Nous avons montré que les estimateurs fondés sur un modèle à fonction p -spline sont généralement plus efficaces que l'estimateur HT et peuvent fournir des intervalles de confiance plus étroits dont la couverture est proche du niveau de confiance nominal. Dans le présent article, nous étendons cette approche aux plans d'échantillonnage à deux degrés. Nous utilisons un modèle mixte fondé sur une p -spline qui est ajusté à une relation non paramétrique entre les moyennes des unités primaires d'échantillonnage (UPF) et une mesure de la taille des UPF, et auquel sont intégrés des effets aléatoires pour modéliser la mise en grappes. Pour l'estimation de la variance, nous considérons les méthodes d'estimation de la variance fondées sur un modèle bayésien empirique, la méthode du jackknife et la méthode des répliques répétées équilibrées (BRR). Des études en simulation portant sur des données simulées et des échantillons tirés des microdonnées à grande diffusion du Recensement de 1990 montrent que l'estimateur fondé sur un modèle à fonction p -spline donne de meilleurs résultats que l'estimateur HT et que les estimateurs linéaires assistés par un modèle. Les simulations montrent aussi que les méthodes d'estimation de la variance produisent des intervalles de confiance dont la couverture est égale à la sélection à la première étape est PPT et où les probabilités de sélection à la deuxième étape sont proportionnelles à l'inverse des probabilités de sélection à la première étape, et où l'estimateur HT mène à la moyenne non pondérée. Dans les situations favorisant le plus l'utilisation de l'estimateur HT, les estimateurs fondés sur un modèle ont une efficacité comparable.

MOTS CLÉS : Pondération; maximum de vraisemblance restreint (REML); estimation bayésienne empirique.

1. INTRODUCTION

Dans une enquête par sondage, soit y_i la valeur d'un résultat X pour l'unité i , et soit S l'ensemble d'unités échantillonnées. L'estimateur d'Horvitz-Thompson (HT) (Horvitz et Thompson 1952) $\hat{Y}_{HT} = \sum_{i \in S} y_i / \pi_i$, où π_i est la probabilité de sélection de l'unité i , est un estimateur sans biais par rapport au plan de sondage du total de population finie (et de la moyenne lorsque on le divise par le chiffre de population connu N). Il peut aussi être considéré comme un estimateur fondé sur un modèle projectif (Firth et Bennett 1998) du modèle linéaire suivant reliant y_i à π_i :

$$y_i = \beta \pi_i + \pi_i \varepsilon_i,$$

où l'on suppose que les ε_i sont i.i.d. et suivent une loi normale de moyenne nulle et de variance σ^2 .

Dans Zheng et Little (2003, 2004), nous avons proposé un modèle non paramétrique

$$y_i = f(\pi_i) + \varepsilon_i, \varepsilon_i \sim \text{ind } N(0, \pi_i^2 \sigma^2),$$

utilisant des splines pénalisées pour modéliser la moyenne du résultat y_i sous forme d'une fonction à variation lisse f des probabilités de sélection π_i . Nous avons montré dans Zheng et Little (2003) que les estimateurs fondés sur un modèle non paramétrique sont plus efficaces que l'estimateur HT pour les échantillons généraux à un degré sélectionnés avec probabilité proportionnelle à la taille (PPT) et ne sont pas nettement moins efficaces que l'estimateur HT lorsque les données sont générées à l'aide d'un modèle favorisant l'estimateur HT.

- POTTER, F. (1990). A study of procedures to identify and trim extreme sampling weights. Dans *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 225-230.
- POTTER, F. (1993). The effect of weight trimming on nonlinear survey estimates. Dans *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 758-763.
- RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, Séries A, 28, 47-60.
- RAO, J.N.K., WU, C.F.J. et YUE, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.
- REN, R., et CHAMBERS, R.L. (2002). Outlier robust imputation of survey data via reverse calibration. Southampton Statistical Sciences Research Institute Methodology, document de travail M03/19, Université de Southampton.
- ROYALL, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.
- ROYALL, R.M., et HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- SÄRNÄL, C.-E., SWENSSON, B. et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New-York, Springer-Verlag.
- STOKES, L. (1990). A comparison of truncation and shrinking of sampling weights. Dans *Proceedings of the 1990 Annual Research Conference*, Washington, DC: Bureau of the Census, 463-471.
- SUGDEN, R.A., et SMITH, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.
- VALLIANT, R., DORFMAN, A. et ROYALL, R.M. (2000). *Finite population sampling: a prediction approach*. New-York, John Wiley & Sons, Inc.
- WEI, A.H., et RONCHETTI, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Séries B*, 60, 413-428.
- ZASLAVSKY, A.M., SCHENKER, N. et BELIN, T.R. (2001). Downweighting influential clusters in surveys: application to the 1990 post enumeration survey. *Journal of the American Statistical Association*, 96, 858-869.

En vertu de l'inégalité de Chebyshev, $t_y / t_x \mathbf{b}$ converge si $t_x \mathbf{b} = O(N)$ et $\sum_{k \in U} \sigma_k^2 = O(N)$ (hypothèse A3). Pour démontrer (b), nous commençons par noter que $E^m E^p(\cdot) = E^p E^m(\cdot | s)$ à condition que l'ensemble de tous les échantillons possibles ne dépende pas de la population qui a été générée par le modèle m . Conséquent, si l'hypothèse (A2) tient, il est simple de montrer que $E^m E^p(t_y / t_x \mathbf{b}) = 1$. Alors, nous notons que

$$V^{mp} \left(\frac{t_y \mathbf{b}}{t_x \mathbf{b}} \right) = V^m E^p \left(\frac{t_y \mathbf{b}}{t_x \mathbf{b}} \right) + E^m V^p \left(\frac{t_y \mathbf{b}}{t_x \mathbf{b}} \right). \quad (A.1)$$

Par conséquent, $V^m E^p(t_y / t_x \mathbf{b}) \leq V^{mp}(t_y / t_x \mathbf{b})$, puisque les deux termes du deuxième membre de (A.1) sont égaux ou supérieurs à 0. En vertu de l'inégalité qui précède et de l'inégalité de Chebyshev, $E^p(t_y / t_x \mathbf{b})$ converge en probabilité vers 1 sous le modèle m quand n et N augmentent, si $\lim_{n, N \rightarrow \infty} V^{mp}(t_y / t_x \mathbf{b}) = 0$. Partant de l'hypothèse (A2), il est simple de montrer que

$$V^{mp} \left(\frac{t_y \mathbf{b}}{t_x \mathbf{b}} \right) = \frac{N}{1} \frac{E^p \left(\frac{t_y \mathbf{b}}{t_x \mathbf{b}} \right)}{1} = \frac{1}{N} \sum_{k \in U} E^p \left\{ \left(\frac{t_y \mathbf{b}}{t_x \mathbf{b}} \right)^2 I_k \right\} \sigma_k^2 / N.$$

Conséquemment, $\lim_{n, N \rightarrow \infty} V^{mp}(t_y / t_x \mathbf{b}) = 0$ si $t_x \mathbf{b} = O(N)$ et $\sum_{k \in U} E^p \left\{ \left(\frac{t_y \mathbf{b}}{t_x \mathbf{b}} \right)^2 I_k \right\} \sigma_k^2 = O(N)$ (hypothèse A3). Ce qui achève la preuve.

BIBLIOGRAPHIE

- BEATON, A.E., et TUKKEY, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147-155.
- BASU, D. (1971). An essay on the logical foundations of survey sampling, part I. Dans *Foundations of statistical inference*, (Éds. V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart, et Winston, 203-233.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.
- CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- CHAMBERS, R.L., DORMAN, A.H. et WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.
- DEVILLE, J.-C., et SÄRNÄDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DRAPER, N., et SMITH, H. (1980). *Applied regression analysis*, second edition. New-York, John Wiley & Sons, Inc.
- POTTER, F. (1988). Survey of procedures to control extreme sampling weights. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453-458.
- PEFFERMANN, D. (1993). The role of sampling weights when modelling survey data. *Revue Internationale de Statistique*, 61, 317-337.
- LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability sampling. *Journal of the American Statistical Association*, 78, 596-604.
- LEE, H. (1995). Outliers in business surveys. *Dans Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott), Chapitre 26, New-York, John Wiley & Sons, Inc.
- LEE, H. (1991). Model-based estimators that are robust to outliers. *Dans Proceedings of the Annual Research Conference*, Washington, DC, U.S. Bureau of the Census, 178-202.
- KORN, E.L., et GRAUBARD, B.L. (1999). *Analysis of Health Surveys*. New-York, John Wiley & Sons, Inc.
- KISH, L. (1992). Weighting for unequal P_j . *Journal of Official Statistics*, 19, 81-97.
- KALTON, G., et FLORES-CERVANTES, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.
- HULLIGER, B. (1999). Simple and robust estimators for sampling. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 54-63.
- HULLIGER, B. (1995). Estimateurs Horvitz-Thompson à l'épreuve des valeurs aberrantes. *Techniques d'Enquête*, 21, 89-97.
- HUBER, P.J. (1981). *Robust Statistics*. New-York, John Wiley & Sons, Inc.
- HUBER, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.
- HEDLIN, D., FALVEY, H., CHAMBERS, R., et KOKIC, P. (2001). Does the model matter for GREG estimation? A business survey example. *Journal of Official Statistics*, 17, 527-544.
- HAMPEL, F.R., RONCHETTI, E.M., ROUSSEURW, P.J., et STAHEL, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. New-York, John Wiley & Sons, Inc.
- GWET, J.-P., et RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.
- GWET, J.-P., et LEE, H. (2000). An evaluation of outlier-resistant procedures in establishment surveys. Dans *The Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria, Virginia, 707-716.
- GWET, J.-P., et LEE, H. (2000). An evaluation of outlier-resistant procedures in establishment surveys. Dans *The Second International Conference on Establishment Surveys*, American Statistical Association, 88, 629-641.
- ELLIOTT, M.R., et LITTLE, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.
- GRAUBARD, B.L., et KORN, E.L. (1993). Hypothesis testing with complex survey data: the use of classical quadratic test statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.
- DUMOUCHEL, W.H., et DUNCAN, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- DUCHESNE, P. (1999). Estimateurs de calage robustes. *Techniques d'enquête*, 25, 47-60.

Tableau 6.2
Comparaison des estimateurs M itérés une seule fois et itérés complètement

Estimateur	Itéré une seule fois	Itéré complètement
Estimateur $M(h_k = 1, \phi = 11)$	BR -6,94 % RREQM -29,28 % ERAM -235,07 %	BR -7,93 % RREQM -29,27 % ERAM -235,07 %
Estimateur $M(h_k = \bar{w}_k, \phi = 100)$	BR -8,14 % RREQM -25,36 % ERAM -197,86 %	BR -8,27 % RREQM -25,33 % ERAM -196,73 %

6.4 Comparaison des estimateurs M itérés une seule fois et itérés complètement

Comparons maintenant les estimateurs M itérés une seule fois et itérés complètement quand $\alpha = 1$. Nous considérons uniquement les deux cas suivants : i) $h_k = 1$ et $\phi = 11$; ii) $h_k = \bar{w}_k$ et $\phi = 100$. En général, l'algorithme IRLS a convergé rapidement dans le cas de l'itération complète (le nombre moyen d'itérations pour obtenir la convergence est de 7,53 pour $h_k = 1$ et de 7,29 pour $h_k = \bar{w}_k$), mais n'a pas convergé pour certains des 5 000 échantillons (64 pour $h_k = 1$ et 75 pour $h_k = \bar{w}_k$). Quand cela s'est produit, nous avons gardé l'estimation M provenant de la dernière itération de l'algorithme IRLS. Il est évident, si l'on examine le tableau 6.2, que le BR, la RREQM et l'ERAM des estimateurs M itérés une seule fois et itérés complètement ont des valeurs avoisinantes. Un point qui mérite d'être souligné est que le BR est un peu plus faible pour les estimateurs M itérés une seule fois, observation qui a été faite par Lee (1991) également et qui est vraisemblablement due au fait que nous avons utilisé $B^{(0)} = B^c$ comme vecteur des valeurs de départ pour l'algorithme IRLS, qui est ASBP pour B.

7. CONCLUSION

Nous considérons, dans le présent article, des estimateurs robustes pour remplacer l'estimateur optimal (BLU). Nous commençons par proposer un compromis entre les estimateurs GREG et BLU, c'est-à-dire l'estimateur des moindres carrés (LS), pour résoudre le problème des écarts par rapport à l'hypothèse du plan ignorables. Nous obtenons l'estimateur LS par rétrécissement des poids de sondage vers la moyenne. En principe, cet estimateur devrait être plus stable que l'estimateur GREG lorsque l'hypothèse du plan ignorables tient approximativement et être moins biaisé que l'estimateur BLU lorsque cette hypothèse n'est pas entièrement satisfaite. Ces suppositions sont confirmées par une étude en simulation portant sur une population créée à partir de données d'enquête réelles. L'estimateur LS offre aussi une certaine protection contre les écarts par rapport aux hypothèses qui sous-tendent le modèle.

Pour résoudre le problème des données aberrantes, nous proposons d'utiliser la technique d'estimation M généralisée pondérée afin de réduire l'influence des unités pour

$$E_m \left(\frac{t_y \mathbf{b}}{t_y} \right) = 1$$

Pour démontrer (a), notons que

$$b) E_p \left(\frac{t_y \mathbf{b}}{t_y \mathbf{b}} \right) \text{ converge en probabilité vers } 1.$$

$$a) E_p \left(\frac{t_y \mathbf{b}}{t_y \mathbf{b}} \right) = t_y / t_y \text{ converge en probabilité vers } 1;$$

il suffit de démontrer que :

Dans cette preuve, nous éliminons le conditionnement sur \mathbf{X} lorsque nous traitons les espérances et les variances par rapport au modèle m , afin de simplifier la notation. En utilisant le théorème de Slutsky, pour montrer que $E_p(t_y - t_y)/t_y$ converge en probabilité vers 0 à mesure que la taille de l'échantillon n et la taille de la population N tendent vers l'infini, sous les hypothèses (A1), (A2) et (A3),

REMERCIEMENTS

Les auteurs remercient sincèrement le rédacteur associé et les trois examinateurs de leurs remarques et suggestions constructives. Ils remercient également Cynthia Bocci et Wesley Yung de leurs commentaires qui ont permis de rendre l'article plus clair.

ANNEXE

de ϕ comprises entre 8 et 20. Dans la région où la RREQM s'approche de sa valeur minimale, l'ERAM est plus faible si $h_k = \tilde{w}_k$. Il semble donc que $h_k = \tilde{w}_k$ pourrait mieux contrôler l'effet des unités influentes que $h_k = 1$. Comme prévu, dans les deux figures, le BR diminue quand ϕ augmente.

Nous avons également considéré les versions pondérées et non pondérées de l'estimateur M obtenues en choisissant adaptativement, pour chaque échantillon sélectionné, la valeur de ϕ qui donne lieu à l'EQQM estimée la plus faible (en utilisant l'équation 5.3) parmi les ensembles de valeurs de ϕ considérés plus haut. La valeur moyenne de ϕ sur les échantillons sélectionnés est de 72,34 pour la version pondérée et de 10,58 pour la version non pondérée. Ces moyennes sont calculées en excluant les échantillons pour lesquels $\phi = \infty$ (13 échantillons pour $h_k = \tilde{w}_k$ et un échantillon pour $h_k = 1$). Les deux moyennes sont proches des valeurs optimales de ϕ observées sur les figures 6.3 et

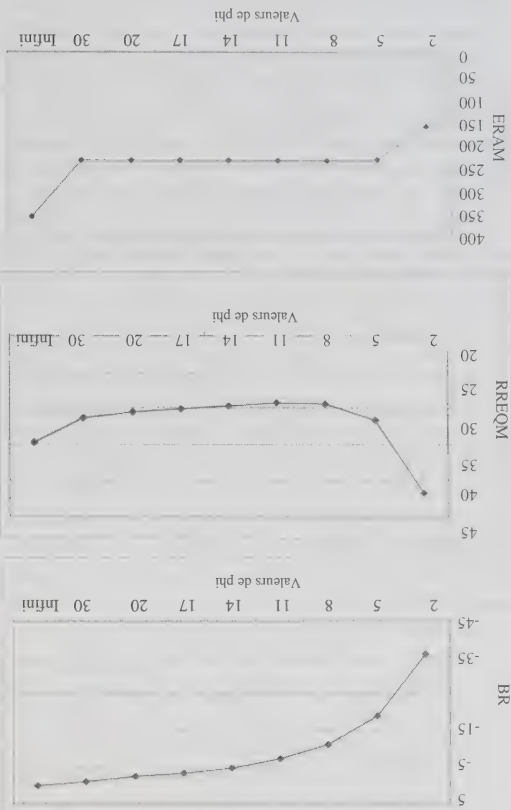


Figure 6.3. BR, RREQM et ERAM de l'estimateur M quand $h_k = 1$

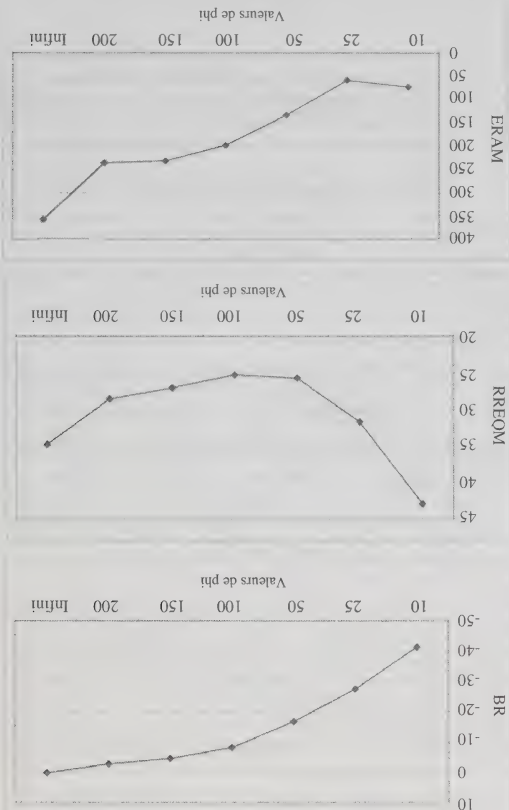
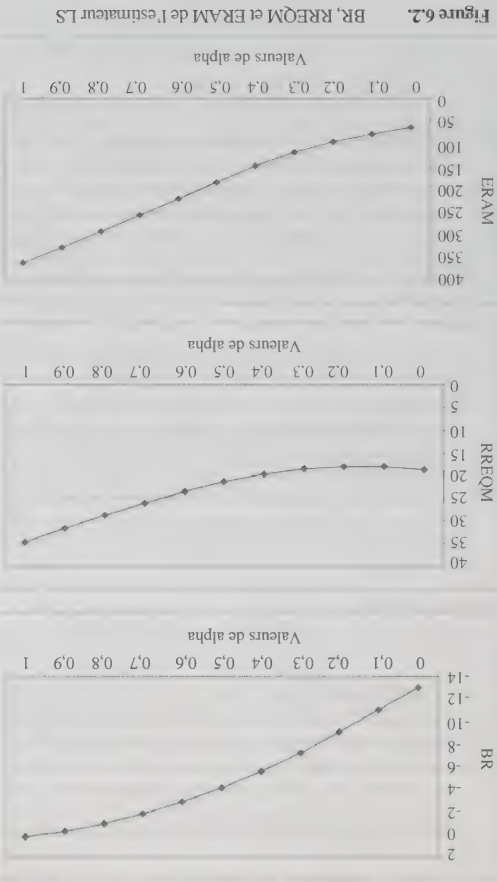


Figure 6.4. BR, RREQM et ERAM de l'estimateur M quand $h_k = \tilde{w}_k$

6.4 (100 pour $h_k = \tilde{w}_k$ et 11 pour $h_k = 1$). La version pondérée de l'estimateur M a un BR de -10,24 %, une RREQM de 28,07 % et une ERAM de 197,86 %. La version non pondérée de l'estimateur M a un BR de -8,26 %, une RREQM de 28,18 % et une ERAM de 232,57 %. Par conséquent, les deux versions de l'estimateur M donnent lieu à une amélioration significative par rapport à l'estimateur GREG en ce qui concerne la RREQM et l'ERAM. L'ERAM est plus faible pour la version pondérée, ce qui, de nouveau, indique que cette dernière contrôle mieux l'effet des unités influentes que la version non pondérée. Cependant, l'écart entre les RREQM des deux estimateurs est très faible. Curieusement, il semble que le fait d'estimer ϕ au lieu d'utiliser la valeur optimale ne fait pas augmenter l'EQQM lorsqu'on utilise la version non pondérée. Cette observation est assez difficile à expliquer.

utilisons un modèle poststratifié, ce résultat donne à penser que l'hypothèse du plan ignorable n'est pas entièrement satisfaisante, même si la corrélation entre les poids de sondage et les résidus de population est faible. D'autre part, l'estimateur GREG ($\alpha=1$) a un très petit BR, mais les RREQM et BRAM les plus grandes, à cause de la variabilité des poids de sondage. Quand $\alpha=0,2$, l'estimateur LS est biaisé, avec un BR de -9,11 %, mais la valeur de son RREQM est la plus faible (17,94 %) parmi les valeurs de α considérées. Il s'agit d'une réduction importante comparativement à la RREQM de l'estimateur GREG (34,77 %). En général, les valeurs de α comprises entre 0,2 et 0,5 donnent un estimateur de compromis raisonnable en ce qui concerne le BR, la RREQM et l'ERAM. Notons que, pour des effectifs d'échantillon prévus plus grands, nous nous attendons à ce que l'EQM minimale soit atteinte pour les valeurs plus grandes de α , parce que le biais de l'estimateur LS peut dominer sa variance.



6.3 Estimateur M : robustesse aux données aberrantes

Nous avons également considéré l'estimateur LS obtenu en choisissant adaptativement, pour chaque échantillon sélectionné, la valeur de α qui donne l'EQM estimée la plus petite parmi l'ensemble des 11 valeurs de α considérées plus haut. Nous avons estimé l'EQM au moyen de l'équation (5.3). La valeur moyenne de α sur les 5 000 échantillons sélectionnés est de 0,43, c'est-à-dire une valeur un peu plus grande que celle de 0 (0,2) que donnant l'EQM la plus petite (voir la figure 6.2). Cette situation pourrait être due à la simplification faite pour obtenir (5.3), c'est-à-dire l'omission d'une composante du carré du biais par rapport au plan de sondage lors de l'estimation de l'EQM. Néanmoins, cet estimateur LS représente une amélioration significative par rapport à l'estimateur GREG en ce qui a trait à la RREQM (26,05 %) et à l'ERAM (217,99 %). Cet estimateur LS donne aussi une amélioration significative par rapport à l'estimateur BLU en ce qui concerne le BR (-6,24 %). Par conséquent, choisir adaptativement la valeur de α semble mener à un compromis utile entre les estimateurs GREG et BLU. Cependant, estimer α au lieu d'utiliser la valeur optimale (quoique inconnue) de α entraîne une pénalité en ce qui concerne la RREQM.

Nous avons comparé deux versions de l'estimateur M, l'une qui réduit l'influence des résidus de population pondérés importants ($h_i = w_i$) et l'autre qui réduit l'influence des résidus de population non pondérés importants ($h_i = 1$). Pour la version pondérée, nous avons choisi sept valeurs de ϕ ($\phi = 10, 25, 50, 100, 150, 200, \infty$) et pour la version non pondérée, nous en avons choisi neuf ($\phi = 2, 5, 8, 11, 14, 17, 20, 30, \infty$). Nous avons considéré uniquement le cas $\alpha = 1$, afin de ne pas confondre l'effet du changement de la valeur de la constante α et celui du changement de la valeur de la constante ϕ . Naturellement, nous pourrions trouver un estimateur plus efficace en choisissant la valeur appropriée de chacune des deux constantes. Il convient de souligner que les résultats sont fondés sur une itération unique de l'algorithme IRLS en utilisant $B^{(0)} = B^c$ comme vecteur des valeurs de départ.

Les figures 6.3 et 6.4 montrent que la version pondérée ($h_i = w_i$) offre de meilleures possibilités de réduire la RREQM et l'ERAM des estimateurs M que la version non pondérée ($h_i = 1$). Les deux graphiques de la RREQM présentent une courbe en forme de U. La courbe de la RREQM pour $h_i = w_i$ montre qu'une valeur de ϕ comprise entre 50 et 150 aboutit à une RREQM comprise entre 25 % et environ 27 %, tandis que la RREQM de l'estimateur GREG (dernier point sur les graphiques) est égale à 34,77 %. La courbe de la RREQM pour $h_i = 1$ montre que la RREQM est d'environ 30 % pour les valeurs

enquête, qui est fondée sur un plan de sondage stratifié à plusieurs degrés, fournit des renseignements sur plusieurs variables recueillis auprès de 15 457 ménages. Nous avons choisi la variable *renovations/repairs* comme variable d'intérêt y , parce qu'elle est plus susceptible que d'autres de présenter des valeurs très grandes. Nous avons créé un vecteur x de trois variables auxiliaires binaires en subdivisant la variable *revenu* en trois catégories (*revenu* ≤ 30 000, 30 000 < *revenu* ≤ 60 000 et *revenu* > 60 000) et nous avons choisi $v_k = 1$, pour tout $k \in U$. Autrement dit, nous avons considéré un modèle d'estimation avec poststratification qui devrait être robuste aux écarts par rapport à l'hypothèse de linéarité. Le coefficient de détermination de population (R^2) est égal à 0,13 pour ce modèle d'estimation. Cette valeur de R^2 est typique pour les enquêtes auprès des ménages.

À partir de cette population, nous avons sélectionné 5 000 échantillons dont l'effectif prévu était de 300 par échantillonnage de Poisson. Nous voulions donner aux ménages des probabilités de sélection assez dispersées afin d'obtenir des poids de sondage variables. Nous avons donc attribué les probabilités de sélection de façon telle qu'elles soient proportionnelles à l'inverse des poids de sondage de l'EDM (poids qui incluent un facteur de correction pour la non-réponse). Les probabilités de sélection sont donc données par $\pi_k = (300 / \sum_{k \in U} \pi_k^*) \pi_k^*$, où π_k^* , pour $k \in U$, est la réciproque du poids de sondage (y compris un facteur de correction pour la non-réponse) provenant des données de l'EDM.

Le tableau 6.1 donne certaines statistiques sommaires pour cette population. Nous notons que les distributions des résidus de population sont très asymétriques et que l'asymétrie s'accroît lorsque les résidus sont multipliés par les poids de sondage. La figure 6.1 montre un graphique des résidus de population en fonction des poids de sondage. Premièrement, nous constatons qu'il existe manifestement une donnée aberrante avec une erreur résiduelle supérieure à 50 000 et un poids de sondage ne s'approchant pas de 1. Heureusement, les poids de sondage les plus extrêmes ne sont pas associés à des résidus de population importants. En outre, bien que le graphique puisse être trompeur à cause du nombre énorme de points qui se chevauchent, il ne semble exister aucune relation évidente entre les résidus de population et les poids de sondage. En fait, le coefficient de corrélation entre les poids de sondage et les résidus de population est de 0,0049. Un coefficient de corrélation aussi faible n'est pas atypique dans le cas des enquêtes auprès des ménages, pour les raisons exposées à la section 3 et donne à penser que l'hypothèse du plan ignorable pourrait tenir approximativement.

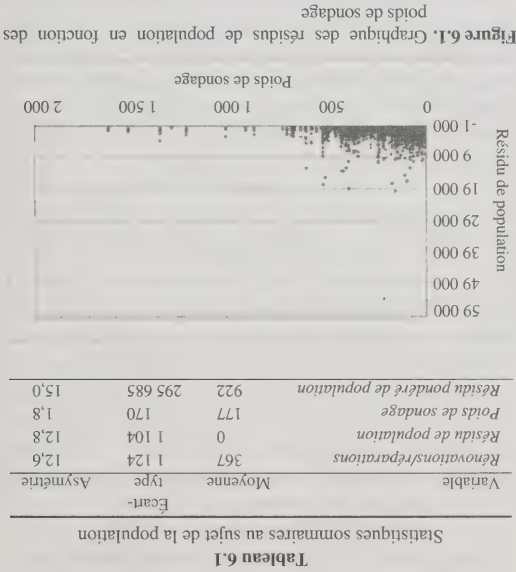


Figure 6.1. Graphique des résidus de population en fonction des poids de sondage

Pour chacun des 5 000 échantillons, nous avons calculé les estimations du total de population pour la variable *renovations/repairs* pour l'estimateur LS, ainsi que pour les deux versions de l'estimateur M , l'une qui réduit l'influence des résidus de population pondérés importants ($h_k = \tilde{w}_k$) et l'autre qui réduit l'influence des résidus de population non pondérés importants ($h_k = 1$). Pour le i^{e} échantillon, l'erreur relative, en pourcentage, de toute estimation \hat{t}_y de t_y est définie comme étant $\Delta_i = 100\% \times (\hat{t}_y - t_y) / t_y$. Le biais relatif (BR) et la racine relative de l'erreur quadratique moyenne (RREQM) de tout estimateur \hat{t}_y , exprimés en pourcentage du total de population, peuvent donc être estimés par $BR = \frac{\sum_{i=1}^{5000} \Delta_i}{5000}$ et $RREQM = \sqrt{\frac{\sum_{i=1}^{5000} \Delta_i^2}{5000}}$, respectivement. Une autre mesure d'intérêt est la valeur absolue maximale de l'erreur relative, ou erreur relative absolue maximale (ERAM), exprimée en pourcentage, donnée par $ERAM = \max\{|\Delta_i|; i = 1, 2, \dots, 5000\}$. Cette mesure peut être utile pour évaluer la sensibilité d'un estimateur à la présence d'unités influentes dans l'échantillon.

6.2 Estimateur LS : robustesse au plan de sondage

À la présente section, nous évaluons les propriétés de l'ERAM de l'estimateur LS pour 11 valeurs de α ($\alpha = 0, 0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9, 1$) quand $g(w_k; \alpha) = w_k^\alpha$. D'une part, l'estimateur BLU ($\alpha = 0$) a une RREQM proche du minimum et l'ERAM la plus faible parmi les 11 valeurs de α , mais, comme prévu, donne lieu à -13,05 %, valeur qui n'est pas négligeable. Comme nous

Les deux derniers termes de (5.1) sont égaux à $[E_p(t_M^y - t_G^y)]^2$. Ils représentent le carré du biais par rapport au plan de t_M^y . Comme l'ont suggéré Gwet et Rivest (1992), un estimateur éventuel de $EQM_p(t_M^y)$ est

$$eqm_p(t_M^y) = \hat{V}_p(t_M^y) + \max(0, (t_M^y - t_G^y)^2 - \hat{V}_p(t_M^y - t_G^y)). \quad (5.2)$$

où $\hat{V}_p(t_M^y)$ et $\hat{V}_p(t_M^y - t_G^y)$ sont les estimateurs de $V_p(t_M^y)$ et $V_p(t_M^y - t_G^y)$, respectivement.

Puisque la structure de l'estimateur t_M^y est complexe, les méthodes d'estimation de la variance par rééchantillonnage offrent un moyen commode d'estimer $V_p(t_M^y)$ et $V_p(t_M^y - t_G^y)$. Les méthodes du jackknife, du bootstrap et des réplics répétées équilibrées (BRR) ont été décrites et évaluées dans Rao, Wu et Yue (1992) pour les plans

d'échantillonnage stratifiés à plusieurs degrés, où il est supposé que les unités primaires d'échantillonnage sont sélectionnées avec remise. Ces auteurs ont montré dans une étude empirique que l'estimateur de la variance par le jackknife peut présenter un biais important lors de l'estimation de la variance d'un estimateur non lisse, comme la médiane d'échantillon. Par conséquent, pour l'estimation de la variance de l'estimateur M , l'estimateur par le

jackknife pourrait être plus biaisé que les estimateurs par les réplics répétées équilibrées ou par le bootstrap si, à chaque itération de l'algorithme IRLS, on estime \hat{Q} au moyen d'un estimateur non lisse tel que (4.4). Gwet et Lee (2000) ont étudié empiriquement les propriétés des méthodes du jackknife et du bootstrap pour certains estimateurs robustes. Dans l'ensemble, leurs résultats sont encourageants. Il importe de souligner que l'estimateur t_M^y doit être recalculé pour chaque rééchantillonnage. Cet exercice inclut la répétition de la procédure utilisée pour estimer α et ϕ si l'on estime ces valeurs au moyen de données d'échantillon courantes.

Quand l'estimation de l'EQM a pour seul but de trouver des valeurs appropriées de α et ϕ , il pourrait être plus commode de considérer des estimateurs simplifiés de l'EQM afin de réduire le temps machine. Aussi proposons-nous quatre moyens de simplifier l'estimation de l'EQM.

1) Nous pourrions nous limiter à une seule itération de l'algorithme IRLS pour chaque rééchantillonnage, même si nous utilisons un estimateur M entièrement itéré. Nous pourrions produire ainsi des estimateurs de la variance raisonnables, puisque les estimateurs M itérés une seule fois et ceux itérés entièrement semblent avoir des propriétés comparables (voir la section 6.4).

ii) Nous pourrions supposer que certaines quantités sont fixes (non aléatoires) pour l'estimation de l'EQM. Cette

hypothèse entraînerait vraisemblablement une sous-estimation de l'EQM, mais l'approche pourrait être utile si l'estimation de l'EQM a uniquement pour but de trouver des valeurs appropriées de α et ϕ . Nous pourrions, par exemple, supposer que les poids corrigés $\tilde{w}_M^k(\hat{B}_M, \hat{Q})$ sont fixes. En fait, cette approximation a été proposée par Hultinger (1999). Autrement, si nous appliquons l'estimateur M en utilisant la méthode décrite à la section 4.2, nous pourrions traiter les valeurs modifiées dans (4.7) ou (4.8) comme des valeurs réelles pour l'estimation de l'EQM.

iii) Nous pourrions omettre le terme $\hat{V}_p(t_M^y - t_G^y)$ dans (5.2), ce qui donnerait l'estimateur de l'EQM suivant : $eqm_p(t_M^y) = \hat{V}_p(t_M^y) + (t_M^y - t_G^y)^2$. Notons que cette approche donne lieu à une surestimation de l'EQM.

iv) Nous pourrions envisager une combinaison de deux des propositions susmentionnées. Par exemple, nous pourrions supposer que les poids corrigés $\tilde{w}_M^k(\hat{B}_M, \hat{Q})$ sont fixes et omettre le terme $\hat{V}_p(t_M^y - t_G^y)$ dans (5.2). Dans ces conditions, nous pourrions obtenir un estimateur de $V_p(t_M^y)$ en notant que $V_p(t_M^y) = (t_M^y - t_G^y)^2$ et en utilisant la technique bien connue de linéarisation par série de Taylor proposée par Binder (1983) pour estimer $V_p(\hat{B}_M)$. Après des calculs algè-

bres simples, nous obtenons l'estimateur de l'EQM

$$eqm_p(t_M^y) = \sum_{k \in les} \frac{\pi_{kI}}{(\pi_{kI} - \pi_k \pi_I)} w_M^k (y^k - x_k^t \hat{B}_M) w_M^t (y^t - x_t^t \hat{B}_M) + (t_M^y - t_G^y)^2, \quad (5.3)$$

où π_{kI} est la probabilité conjointe de sélection des unités k et I .

6. ÉTUDE EN SIMULATION

Nous avons réalisé une étude en simulation pour évaluer certaines propriétés de l'estimateur LS et de l'estimateur M pour une population finie asymétrique. Plus précisément, nous avons comparé une version de l'estimateur M qui réduit l'influence des résidus pondérés de population dont la valeur est grande à une autre qui réduit l'influence des grands résidus non pondérés de population. Nous avons également comparé les propriétés des estimateurs M itérés une seule fois et itérés entièrement. La section 6.1 décrit la population et le plan de sondage, et les sections 6.2 à 6.4, les résultats de la simulation.

6.1 Population et plan de sondage

Nous avons utilisé comme population les données (EDM) de 1998 réalisées par Statistique Canada. Cette

variables différentes. Pour simplifier la notation, nous représentons les poids corrigés associés à la variable y_i par $\tilde{w}_i^k(y_i)$, pour $i = 1, 2, \dots, q$. Puisque les poids corrigés $\tilde{w}_i^k(y_i)$ dépendent de la variable d'intérêt y_i , nous obtenons q ensembles de poids, même si nous choisissons une valeur commune de ϕ .

Gwet et Rivest (1992), Duchesne (1999) et Hülliger (1999) ont proposé d'utiliser les poids corrigés $\tilde{w}_i^k(y) = \min(\tilde{w}_i^k(y_1), \tilde{w}_i^k(y_2), \dots, \tilde{w}_i^k(y_q))$ pour obtenir un ensemble unique d'estimation $\tilde{w}_i^k(y)$ pour $\tilde{w}_i^k(y)$ dans (4.5) et calculés en remplaçant $\tilde{w}_i^k(\mathbf{B}_M, \hat{Q})$ par $\tilde{w}_i^k(y)$. Bien que les poids d'estimation $\tilde{w}_i^k(y)$ soient calés sur les totaux connus de population \mathbf{t}_x , ils ne sont pas calés sur le vecteur des estimations \mathbf{t}_M^k , que nous considérons comme nos meilleures estimations au sens de la minimisation de l'EQM estimée. De surcroît, l'utilisation de $\sum_{k \in s} \tilde{w}_i^k(y) y_k$ produit vraisemblablement un biais par rapport au plan de sondage plus important que \mathbf{t}_M^k , même si elle permet de contrôler la variance par rapport au plan de sondage. Pour résoudre ces problèmes, nous proposons de calculer les poids d'estimation $\tilde{w}_i^k(\mathbf{B}_M, \hat{Q})$ en remplaçant $\tilde{w}_i^k(y)$ par les poids corrigés $\tilde{w}_i^k(y)$ dans (4.5) et en augmentant le vecteur de variables auxiliaires \mathbf{x} et les totaux connus de population \mathbf{t}_x au moyen de y et \mathbf{t}_M^k , respectivement. Par conséquent, les poids d'estimation $\tilde{w}_i^k(\mathbf{B}_M, \hat{Q})$ sont calés sur \mathbf{t}_x et \mathbf{t}_M^k , et \mathbf{t}_M^k est estimé au moyen de $\mathbf{t}_M^k = \sum_{k \in s} \tilde{w}_i^k(\mathbf{B}_M, \hat{Q}) y_k$. Naturellement, on pourrait vouloir limiter le nombre de variables de calage, ce qui restreindrait dans une certaine mesure l'applicabilité de cette méthode lorsque la valeur de q est très grande.

4.2 Modification des valeurs des unités influentes

Un autre moyen d'appliquer l'estimateur M, \mathbf{t}_M^k , en pratique consiste à modifier les valeurs des variables d'intérêt y et à utiliser les poids d'estimation \mathbf{t}_M^k , pour toutes les variables. Nous pouvons faire cela séparément pour chaque variable d'intérêt, si bien que nous revenons à la présente section au cas d'une variable d'intérêt unique. Commentons par représenter par s_o l'ensemble aléatoire de toutes les unités échantillonnées k pour lesquelles $\tilde{w}_i^k(\mathbf{B}_M, \hat{Q}) \neq \tilde{w}_i^k$. Autrement dit, s_o est l'ensemble aléatoire d'unités qui ont été reconnues comme étant influentes. Posons aussi que \mathbf{B}_{M^*} est défini implicitement par l'équation

$$\sum_{k \in s} \tilde{w}_i^k(y_k - x_k \mathbf{B}_{M^*}) \frac{y_k}{x_k} = 0, \quad (4.6)$$

où $y_k = y_i$, si $k \in s - s_o$, et $y_k = y_k^*$, autrement. La quantité y_k^* est une valeur modifiée de l'unité influente k que nous utilisons pour remplacer y_k . Notons que

$\mathbf{B}_{M^*} = \mathbf{B}_{LS}$ si $y_k = y_i$, pour $k \in s$. Nous pouvons alors estimer le total de population $t_{y_{M^*}}^k$ par $t_{y_{M^*}}^k = \mathbf{t}_M^k$. Il est également facile de montrer que $t_{y_{M^*}}^k = \sum_{k \in s} \tilde{w}_i^k y_k^*$. L'idée ici consiste à trouver des valeurs modifiées y_k^* , pour $k \in s_o$, aussi proches que possible des valeurs originales y_k et satisfaisant à la contrainte $\mathbf{B}_{M^*} = \mathbf{B}_M$. Sous cette contrainte, il est évident que $t_{y_{M^*}}^k = t_{y_M}^k$. Nous obtenons une application de cette idée en minimisant la fonction de distance $\sum_{k \in s_o} \tilde{w}_i^k (y_k^* - y_k)^2 / y_k$ sous la contrainte $\mathbf{B}_{M^*} = \mathbf{B}_M$. Cela nous mène aux valeurs modifiées

$$y_k^* = y_k + \mathbf{x}_k' \left(\sum_{k \in s_o} \tilde{w}_i^k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in s} \tilde{w}_i^k \mathbf{x}_k \mathbf{x}_k' \right) (\mathbf{B}_M - \hat{\mathbf{B}}_{LS}). \quad (4.7)$$

Cette idée est essentiellement équivalente au calage inverse proposé par Ren et Chambers (2002), excepté que ces auteurs ont utilisé la contrainte $t_{y_{M^*}}^k = t_{y_M}^k$ au lieu de $\mathbf{B}_{M^*} = \mathbf{B}_M$. Par exemple, il est facile de montrer que cette des valeurs modifiées qui satisfont à la contrainte $\mathbf{B}_{M^*} = \mathbf{B}_M$. Par conséquent, cette valeur observée est fortement influente.

où $a_k = \tilde{w}_i^k(\mathbf{B}_M, \hat{Q}) / \tilde{w}_i^k$. Les valeurs modifiées de l'équation (4.8) ont une interprétation simple : il s'agit de la moyenne pondérée de la prédiction robuste $\mathbf{x}_k' \mathbf{B}_M$ et de la valeur observée y_k . Moins de poids est accordé à la valeur observée y_k quand la valeur de a_k est plus faible et que, par conséquent, cette valeur observée est fortement

5. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE

L'estimation de l'EQM de $t_{y_M}^k$ peut avoir trois objectifs : i) trouver des valeurs apprises de α et ϕ au moyen de données d'échantillon anciennes ou courantes, ii) évaluer la qualité des estimations et iii) faire des inférences au sujet de quantités de population inconnues. Partant du fait que quantités de population inconnues, il est facile de montrer que l'EQM de $t_{y_M}^k$ peut être approximée par

$$\text{EQM}_p(t_{y_M}^k) \approx V_p(t_{y_M}^k) + E_p(t_{y_M}^k - t_{y_o}^k)^2 - V_p(t_{y_M}^k - t_{y_o}^k). \quad (5.1)$$

Une meilleure approche pourrait consister à tronquer les résidus pondérés dont la valeur est grande. Nous pouvons pour cela utiliser la technique de l'estimation M généralisée pondérée décrite à la section suivante.

L'estimateur LS_{γ}^{γ} peut aussi s'écrire sous la forme $t_{\gamma}^{\gamma} = \sum_{k \in s} w_k^{\gamma} y_k$, où les poids d'estimation LS sont

$$w_k^{\gamma} = \frac{x_k^{\gamma}}{\sum_{k \in s} \frac{x_k^{\gamma}}{w_k^{\gamma}}} \quad (3.6)$$

Notons que les poids d'estimation w_k^{γ} , y compris w_k^{γ} et $w_k^{\gamma} x_k^{\gamma}$ en tant que cas particuliers, sont calés sur les totaux

w_c^{γ} connus de population t_{γ}^{γ} en ce sens qu'ils satisfont l'équation de calage $\sum_{k \in s} w_k^{\gamma} x_k^{\gamma} = t_{\gamma}^{\gamma}$ (voir Deville et Samdal 1992).

4. ROBUSTESSE AU MODÈLE (AUX DONNÉES ABRÉVANTES)

Comme nous le mentionnons dans l'introduction,

l'estimateur LS_{γ}^{γ} protégé dans une certaine mesure contre les écarts par rapport à l'hypothèse de plan ignorable,

ainsi que contre les écarts par rapport aux hypothèses qui sous-tendent le modèle. Cependant, il n'offre aucune

protection contre les données aberrantes, qui peuvent être considérées comme la conséquence d'une erreur de

spécification du modèle, y compris un écart par rapport à l'hypothèse de normalité. Ainsi, l'estimateur GREG est

ASBP quelle que soit la validité du modèle d'estimation. Toutefois, sa variance par rapport au plan peut être

grande en présence de données aberrantes dans la population finie, parce qu'elles peuvent influencer fortement sur son

erreur d'échantillonnage lorsqu'elles sont sélectionnées dans l'échantillon. Ce problème risque d'être amplifié lorsque les

poids de sondage sont très dispersés. Dans le cas de l'estimateur d'Horvitz-Thompson, cette situation a été bien

illustrée par l'exemple du cirque de Basu (1971). Naturellement, l'utilisation de variables auxiliaires efficaces

au stade de l'estimation peut permettre de neutraliser l'effet des valeurs aberrantes sur les estimations. Malheureusement, très souvent, ce genre de variables auxiliaires ne sont

pas disponibles, si bien que l'utilisation d'estimateurs robustes aux données aberrantes peut donner des résultats

significativement meilleurs que l'estimateur LS .

En utilisant la technique de linéarisation par série de Taylor (voir, par exemple, Samdal et coll. 1992, page 235) et sachant que $t_{\gamma}^{\gamma} = t_{\gamma}^{\gamma} B$, il est bien connu, et facile de montrer, qu'on peut approximer l'erreur d'échantillonnage de l'estimateur GREG comme suit : $t_{\gamma}^{\gamma} - t_{\gamma}^{\gamma} \approx \sum_{k \in s} w_k^{\gamma} E_k$, où $E_k = y_k - x_k^{\gamma} B$ est le résidu de population pour l'unité k . Par conséquent, un poids de sondage élevé associé à un résidu de population (ou valeur aberrante) important peut avoir un effet considérable sur la qualité de l'estimateur

$$\sum_{k \in s} w_k^{\gamma} \frac{1}{h_k} \psi \left(\frac{h_k \tilde{E}_k(\mathbf{B}_M)}{x_k^{\gamma}} \right) = 0, \quad (4.1)$$

L'estimateur M de \mathbf{B}_M de \mathbf{B} , qui est défini implicitement par

pondéré de population est grand. Nous obtenons ainsi pondérée pour réduire l'influence des unités dont le résidu pages 315 – 316) de la technique d'estimation M généralisée utilisons la version de Schweppe (Hampel et coll. 1986, qualité de l'estimateur LS . Pour résoudre ce problème, nous peut influencer fortement l'erreur d'échantillonnage et la mation élevée associée à un résidu de population important la forme $t_{\gamma}^{\gamma} - t_{\gamma}^{\gamma} = \sum_{k \in s} w_k^{\gamma} E_k$. Donc, un poids d'esti- d'échantillonnage de l'estimateur LS peut s'exprimer sous GREG. De surcroît, il est facile de montrer que l'erreur

régression linéaire pondérée :

$$\sum_{k \in s} w_k^{\gamma} (\mathbf{B}_M, \tilde{O}) (y_k - x_k^{\gamma} \mathbf{B}_M) / x_k^{\gamma} = 0, \quad (4.2)$$

L'équation (4.1) peut s'écrire sous la forme d'une aberrantes doivent exercer une certaine influence sur \mathbf{B}_M .

dans la population finie. Par conséquent, les données fortement influencé par la présence de données aberrantes robuste (au sens classique) pour \mathbf{B} , puisqu'il peut être des paramètres du modèle \mathbf{B} . En fait, \mathbf{B} lui-même n'est pas vecteur des paramètres de population \mathbf{B} et non du vecteur nouveau que nous recherchons un estimateur robuste du résidus non pondérés. En outre, il convient de souligner de $h_k = \sqrt{V_k}$ ou $h_k = 1$, qui réduisent l'influence des grands et que les choix proposés plus haut de vraiement tous deux donner de meilleurs résultats que les choix plus simples qui ne tiennent pas compte des variables auxiliaires \mathbf{z} , tels que h_k devrait dépendre des poids de sondage w_k^{γ} ou w_k^{γ} supposant que $g(w_k^{\gamma}; \alpha) = w_k^{\gamma}$. Le point important ici est auxiliaires \mathbf{x} ou que la valeur de α n'est pas proche de 1 (en s'il existe des valeurs aberrantes dans les variables pourrait choisir $h_k = w_k^{\gamma} \sqrt{V_k}$ de préférence à $h_k = w_k^{\gamma} \sqrt{V_k}$ dans le second, celle d'une grande valeur de $w_k^{\gamma} E_k$. On l'influence d'une grande valeur de $w_k^{\gamma} E_k$ qui est réduite et $\sqrt{V_k}$ est un choix naturel. Dans le premier cas, c'est considérations qui précèdent, $h_k = w_k^{\gamma} \sqrt{V_k}$ ou $h_k =$ pour lesquelles $h_k E_k(\mathbf{B})$ est grand. Compte tenu des de la fonction $\psi(\cdot)$ consiste à réduire l'influence des unités dépendre non seulement de \mathbf{x}_k , mais aussi de \mathbf{z}_k . Le rôle d'échelle de population positif et h_k est un poids qui peut où $\tilde{E}_k(\mathbf{B}_M) = (y_k - x_k^{\gamma} \mathbf{B}_M) / \sqrt{V_k}$, \tilde{O} est un paramètre

où

$$w_k^{\gamma} \mathbf{B}_M, \tilde{O} = w_k^{\gamma} \psi(t_{\gamma}^{\gamma})$$

soient satisfaites entièrement. Par exemple, nous pourrions hésiter à inclure tous les identifiants de strate dans le modèle d'estimation si le nombre de strates est très grand. Le cas échéant, l'hypothèse de plan ignorable pourrait ne pas être entièrement vérifiée. En outre, le modèle d'estimation, y compris l'hypothèse de normalité, pourrait ne pas être vérifié pour chaque variable d'intérêt. Conséquence, il n'est pas toujours approprié d'utiliser aveuglément l'estimateur BLU t_y^B de t_y , et il faut parfois recourir à des estimateurs robustes.

3. ROBUSTESSE AU PLAN DE SONDAGE

Partant du fait que $v_k = \mathbf{x}_k^* \lambda$, il est facile de montrer (voir Särndal, Swensson et Wretman, 1992 page 231) que t_y peut s'exprimer sous la forme $t_y = \mathbf{t}_y^* \mathbf{B}$, où \mathbf{B} est défini implicitement par l'équation

$$\sum_{k \in S} (y_k - \mathbf{x}_k^* \mathbf{B}) \frac{v_k}{\mathbf{x}_k^*} = 0. \quad (3.1)$$

Le vecteur \mathbf{B} serait l'estimateur LS de β , sous le modèle d'estimation m , si l'on pouvait réaliser un recensement. Puisque \mathbf{t}_k est connu, rechercher un estimateur du total de la théorie de l'estimation fondée sur le plan de sondage, un estimateur naturel \mathbf{B}_G^* de \mathbf{B} est défini implicitement par l'équation

$$\sum_{k \in S} v_k (y_k - \mathbf{x}_k^* \mathbf{B}_G^*) \frac{v_k}{\mathbf{x}_k^*} = 0, \quad (3.2)$$

où w_k , le poids de sondage de l'unité k , est égal à l'inverse de la probabilité de sélection π_k . L'utilisation de \mathbf{B}_G^* mène à l'estimateur GREG $t_y^G = \mathbf{t}_y^* \mathbf{B}_G^*$ de t_y . L'estimateur GREG t_y^G prend une forme de projection simple parce que $v_k = \mathbf{x}_k^* \lambda$ (voir Särndal et coll. 1992, page 231). Nous pouvons aussi l'écrire sous la forme $t_y^G = \sum_{k \in S} w_k^G y_k$, où les poids d'estimation GREG w_k^G sont donnés par

$$w_k^G = w_k \frac{\mathbf{x}_k^*}{\mathbf{x}_k^* \mathbf{x}_k^*} \left(\sum_{k \in S} w_k \frac{v_k}{\mathbf{x}_k^* \mathbf{x}_k^*} \right)^{-1} \mathbf{t}_k. \quad (3.3)$$

Comme nous le soulignons dans l'introduction, l'estimateur GREG est robuste au biais puisque sa propriété d'être ASBP ne dépend pas de la validité du modèle d'estimation m ni de l'hypothèse de plan ignorable. Cependant, l'estimateur GREG n'est pas robuste à la variance, puisqu'il est parfois nettement moins efficace que l'estimateur BLU lorsque les deux hypothèses sont vérifiées. L'inefficacité de l'estimateur GREG est due à la forte dispersion des poids de sondage. Dans le cas des enquêtes auprès des ménages, cette situation n'est pas rare, à cause du calage. En outre, les considérations pratiques quant au choix

$$\sum_{k \in S} \tilde{w}_k (y_k - \mathbf{x}_k^* \mathbf{B}_{LS}) \frac{v_k}{\mathbf{x}_k^*} = 0 \quad (3.4)$$

et où \tilde{w}_k est le poids « rétréci » de l'unité k donné par

$$\tilde{w}_k = \frac{\sum_{k \in S} w_k}{\sum_{k \in S} g(w_k; \alpha)} g(w_k; \alpha), \quad (3.5)$$

Dans le deuxième membre de (3.5), le rôle du ratio est simplement d'assurer que $\sum_{k \in S} \tilde{w}_k = \sum_{k \in S} w_k$ et celui de la fonction $g(w_k; \alpha)$, d'obtenir les poids rétrécis \tilde{w}_k qui sont moins variables que les poids de sondage w_k . Nous supposons que cette fonction est monotone en la constante α , avec $1 \leq g(w_k; \alpha) \leq w_k$. Les estimateurs BLU et GREG sont par conséquent des cas spéciaux extrêmes de l'estimateur LS obtenus quand α est tel que $g(w_k; \alpha) = 1$ et $g(w_k; \alpha) = w_k$, respectivement. Pour obtenir un simple compromis entre ces deux estimateurs extrêmes, nous proposons d'utiliser $g(w_k; \alpha) = w_k^\alpha$, avec $0 \leq \alpha \leq 1$. Le choix $\alpha = 0$ mène à l'estimateur BLU, tandis que le choix $\alpha = 1$ mène à l'estimateur GREG. En fait, l'auteur de cette proposition est Kish (1992, page 198). D'autres fonctions $g(w_k; \alpha)$ et d'autres moyens de réduire la variabilité des poids de sondage sont décrits dans la littérature (voir, par exemple, Elliott et Little 2000). La troncation des poids de sondage de valeur élevée ($g(w_k; \alpha) = \min(w_k; \alpha)$, avec $\alpha > 0$) est une approche fréquemment utilisée pour résoudre ce problème. Elle peut être utile lorsque les hypothèses (A_1) et (A_2) ne sont pas entièrement satisfaites et qu'il existe des poids de sondage anormalement grands.

du plan de sondage, conjuguées à l'information limitée dont on dispose à l'étape de la conception du plan de sondage, produisent fréquemment des plans de sondage approximatifs et ignorables. Ainsi, dans le cas des enquêtes auprès des ménages, l'information géographique est souvent la principale information auxiliaire dont on dispose pour construire les strates. À moins que le nombre de strates soit très grand, cette information est, en général, faiblement corrélée aux variables quantitatives d'intérêt, telles que les dépenses ou le revenu, et à leur variable d'erreur résiduelle correspondante $E = y - \mathbf{x}^* \mathbf{B}$. Par conséquent, la variabilité de poids de sondage w est, elle aussi, séquentiellement corrélée à E , ce qui donne à penser que l'hypothèse de plan ignorable pourrait tenir approximativement. Cela fait aussi penser que les poids de sondage agissent plus ou moins comme un bruit aléatoire lors de l'estimation de \mathbf{B} au moyen de (3.2) et que leur influence pourrait être significativement réduite. Par conséquent, pour obtenir un estimateur robuste au plan de sondage lorsque la variabilité des poids de sondage est forte, nous proposons de rétrécir (*shrink*) les poids de sondage de façon qu'ils se rapprochent de la moyenne et d'utiliser l'estimateur LS $t_y^{LS} = \mathbf{t}_y^* \mathbf{B}_{LS}$, où \mathbf{B}_{LS} est défini implicitement par

rapport à l'hypothèse de linéarité, nous pouvons utiliser un modèle poststratifié quand il est possible de partitionner la population en groupes homogènes et mutuellement exclusifs. Hedlin, Falvey, Chambers et Kocio (2001) illustrent l'importance d'une modélisation minutieuse dans le cas des enquêtes par sondage.

L'hypothèse (A2) est une condition suffisante pour que l'on puisse ignorer (Rubin 1976) le plan de sondage en ce qui concerne la loi de y sachant \mathbf{X} . Autrement dit, la loi de y est indépendante de s après le conditionnement sur \mathbf{X} . En utilisant l'hypothèse (A1), nous pouvons subdiviser y en un terme fixe $\mathbf{X}\beta$ et un terme d'erreur aléatoire $\varepsilon = y - \mathbf{X}\beta$. Par conséquent, si le plan de sondage est indépendant de ε sachant \mathbf{X} , c'est-à-dire si $p(s|\mathbf{X}) = p(s|\varepsilon, \mathbf{X})$, alors l'hypothèse (A2) est satisfait et le plan de sondage est ignorable. Puisque nous considérons uniquement les plans de sondage de la forme $p(s|\mathbf{Z})$, un moyen évident de rendre le plan de sondage ignorable consiste à inclure toutes les variables de plan de sondage \mathbf{z} dans le modèle d'estimation. Les variables utilisées pour former les strates, celles utilisées comme mesure de taille en cas d'échantillonnage avec probabilité proportionnelle à la taille, et ainsi de suite, sont des exemples de variables de plan de sondage. Les poids de sondage peuvent aussi résumer utilement l'information sur le plan de sondage. Notons qu'il ne faut pas nécessairement inclure toutes les variables de plan de sondage dans le modèle d'estimation (voir Sugden et Smith 1984). Les variables de plan qui sont indépendantes de y (ou ε) sachant \mathbf{X} ne doivent pas être incluses dans le modèle. Pour évaluer la validité de l'hypothèse (A2), l'utilisation d'un graphique des résidus $\varepsilon_i = y_i - \mathbf{x}_i^k \beta$ en fonction des poids de sondage w_i (ou de toute variable de plan) pourrait être utile (voir Pfeffermann 1993). Toute tendance se dégageant de ce graphique laisse entendre que les poids de sondage sont corrélés à l'erreur aléatoire ε et que le plan de sondage n'est pas ignorable en ce qui concerne le modèle d'estimation. On peut aussi recourir à des tests plus formels pour évaluer la validité de cette hypothèse (voir, par exemple, DuMouchet et Duncan 1983; Graubard et Korn 1993; et, pour d'autres références bibliographiques à ce sujet, Pfeffermann 1993).

Sous le modèle d'estimation m et l'hypothèse de plan ignorable (A2), il est facile de montrer que l'estimateur BLU (Royall 1976) \hat{t}_y^B de t_y prend la forme de projection simple $\hat{t}_y^B = \mathbf{t}_y^B \mathbf{B}^B$, où \mathbf{B}^B est défini implicitement par l'équation

$$\sum_{k \in s} (y_k - \mathbf{x}_k^B \mathbf{B}^B) \frac{V_k}{\mathbf{x}_k^B} = 0. \quad (2.1)$$

L'estimateur BLU peut aussi s'écrire sous la forme $\hat{t}_y^B = \sum_{k \in s} w_k^B y_k$, où les poids d'estimation BLU w_k^B sont donnés par

$$w_k^B = \frac{\mathbf{x}_k^B}{\sum_{k \in s} \mathbf{x}_k^B \mathbf{x}_k^B} \left(\mathbf{t}_k^B \right)^{-1} \quad (2.2)$$

La variance par rapport au modèle $V_m(t_y^B | s, \mathbf{X})$ de \hat{t}_y^B est celle qui est la plus petite pour chaque échantillon possible parmi tous les estimateurs linéaires sans biais par rapport à m de t_y . Une conséquence directe de ce résultat est que la variance anticipée $E_m(E_p(t_y^B | \mathbf{X})^2 | \mathbf{X})$ est également la plus petite parmi tous les estimateurs linéaires sans biais par rapport à m de t_y , où l'indice inférieur p indique que l'espérance est évaluée par rapport au plan de sondage. Sous l'hypothèse supplémentaire que la loi de y_k , sachant \mathbf{X} , est normale, \mathbf{B}^B est aussi l'estimateur du maximum de vraisemblance du vecteur de paramètres du modèle β .

En général, l'estimateur BLU \hat{t}_y^B n'est pas asymptotiquement sans biais par rapport au plan (ASBP). Cependant, sous le modèle d'estimation m , l'hypothèse de plan ignorable (A2) et l'hypothèse supplémentaire (A3) énoncée plus loin, l'estimateur BLU a la propriété d'être asymptotiquement sans biais par rapport au plan en probabilité (ASBPP) en ce sens que son biais relatif par rapport au plan $(\hat{t}_y^B - t_y)/t_y$ converge en probabilité vers 0 quand n et N augmentent sans borne.

(A3) $\sum_{k \in U} E_p\{w_k^B I_k | \sigma_k^2 = O(N)\}$, $\sum_{k \in U} \mathbf{x}_k^B \beta = O(N)$ et $\sum_{k \in U} \sigma_k^2 = O(N)$, où $\sigma_k^2 = \sigma^2 V_k$ et I_k est une variable aléatoire fictive indiquant si l'unité k est sélectionnée dans l'échantillon ($I_k = 1$) ou ne l'est pas ($I_k = 0$).

L'hypothèse (A3) décrit le comportement asymptotique de trois quantités de population. Plus précisément, exiger que $\sum_{k \in U} E_p\{w_k^B I_k | \sigma_k^2 = O(N)\}$ signifie essentiellement qu'aucun poids d'estimation BLU ne devienne trop grand à mesure que la taille d'échantillon et la taille de population augmentent. Par exemple, si $\mathbf{x}_k = V_k = 1$ et que nous utilisons un plan d'échantillonnage de taille fixe n , alors la condition $\sum_{k \in U} E_p\{w_k^B I_k | \sigma_k^2 = O(N)\}$ équivaut à supposer que les poids $w_k^B = N/n$ restent bornés quand n et N augmentent l'un et l'autre. La preuve que \hat{t}_y^B est ASBPP est donnée en annexe et n'exige pas que $V_k = \mathbf{x}_k^B \lambda$. Par conséquent, l'estimateur BLU est ASBPP, même si la variance par rapport au modèle $V_m(y_k | \mathbf{X})$ est spécifiquement incorrectement.

Comme nous l'avons fait remarquer plus haut, l'estimateur BLU est efficace si le modèle d'estimation m et l'hypothèse de normalité tiennent, de même que l'hypothèse de plan ignorable (A2). Sous ces hypothèses et l'hypothèse supplémentaire (A3), l'estimateur BLU est également ASBPP. Donc, une première étape vers la robustesse consiste à sélectionner et à valider un modèle d'estimation tel que ces hypothèses soient satisfaites dans la mesure du possible. Malheureusement, en pratique, il est rare qu'elles

finies que pour les populations infinies. Comme le fait remarquer Chambers (1986), dans le cas des populations finies, c'est l'erreur d'échantillonnage d'un estimateur (ou l'erreur de prédiction dans un cadre fondé sur un modèle), et non pas nécessairement l'estimateur proprement dit, qui doit être insensible aux données aberrantes. Par exemple, lorsqu'on utilise un plan d'échantillonnage aléatoire simple, la médiane d'échantillon est robuste au sens classique. Par conséquent, la présence d'une valeur aberrante dans la population finie n'a pour ainsi dire aucun effet sur sa variance par rapport au plan de sondage, aussi grande que soit la valeur aberrante. Par contre, l'erreur d'échantillonnage et le biais par rapport au plan de sondage, prennent une valeur arbitrairement grande lorsqu'une ou plusieurs unités de la population prennent une valeur arbitrairement grande, de la population prennent une valeur arbitrairement grande, Cette situation est due au fait que la moyenne de population finie prend, elle-même, une valeur arbitrairement grande dans de telles conditions. Contrairement à la médiane d'échantillon, la moyenne d'échantillon est sans biais par rapport au plan de sondage, mais elle n'est pas robuste au sens classique. La présence d'une donnée aberrante dans la population finie peut donc avoir un effet important sur l'erreur d'échantillonnage et la variance par rapport au plan de sondage de la moyenne d'échantillon. Cela explique pourquoi, dans le cas de populations finies, on considère souvent la robustesse aux données aberrantes comme étant un compromis entre le biais et la variance et pourquoi les données aberrantes doivent habituellement avoir une influence, du moins dans une certaine mesure, sur les estimateurs. L'erreur quadratique moyenne (EQM) est par conséquent un critère utile pour évaluer la qualité des estimateurs des paramètres de population finie robustes aux données aberrantes.

L'objectif réel du présent article est de trouver un estimateur robuste pour remplacer l'estimateur GREG habituellement utilisé de t_y . Cependant, quand on discute de questions de robustesse, il est plus naturel de commencer par introduire l'estimateur optimal (BLU). Donc, à la section 2, nous discutons des hypothèses qui sous-tendent cet estimateur. Nous y donnons aussi les conditions supplémentaires sous lesquelles le biais asymptotique par rapport au plan de sondage de l'estimateur BLU est négligeable. À la section 3, nous traitons de la robustesse au plan de sondage et introduisons l'estimateur des moindres carrés (LS) pondéré. À la section 4, nous discutons de la robustesse au modèle (plus précisément, de la robustesse aux données aberrantes) et nous proposons la technique d'estimation M généralisée pondérée pour réduire l'influence des unités dont le résidu pondéré de population est grand. L'estimateur proposé est convergent dans le cas d'un grand écart par rapport à m β de β . Toute tendance qui se dégage de ce graphique est une indication que la relation entre y et x n'est pas linéaire. Pour obtenir la robustesse aux écarts par

A1) Le modèle d'estimation m qui suit tient : les y_k sachant X , pour $k \in U$, suivent des lois indépendantes de moyenne $E^m(y_k | X) = x_k^T \beta$ et de variance $V^m(y_k | X) = \sigma^2 v_k$, où σ^2 sont des paramètres inconnus du modèle, $v_k = x_k^T \lambda$ et λ est un vecteur de constantes connues. L'indice inférieur « m » indique que les espérances et les variances sont évaluées par rapport au modèle m .

A2) Le plan de sondage est indépendant de y sachant X ; autrement dit, $p(s | y, X) = p(s | X)$, où y est un vecteur contenant N éléments dont le k^{e} est égal à y_k .

L'hypothèse (A1) décrit le modèle d'estimation m , qui spécifie la loi de y sachant X . Nous pouvons utiliser des techniques standard pour valider ce modèle (voir, par exemple, Draper et Smith 1980, chapitre 3). L'hypothèse de linéarité $E^m(y_k | X) = x_k^T \beta$ est une importante hypothèse sous-jacente au modèle d'estimation m . Il existe plusieurs moyens d'en évaluer la validité. On propose souvent, pour cela, d'utiliser un graphique des résidus $e_k = y_k - x_k^T \beta$ en fonction de $x_k^T \beta$ pour un certain estimateur sans biais par rapport à m β de β . Toute tendance qui se dégage de ce graphique est une indication que la relation entre y et x n'est pas linéaire. Pour obtenir la robustesse aux écarts par

2. MEILLEUR ESTIMATEUR LINÉAIRE SANS BIAIS

recensement en ce sens qu'il est égal au total de population finie t_y lorsqu'on réalise un recensement. Nous proposons deux moyens pratiques d'appliquer les estimateurs M aux enquêtes polyvalentes; soit nous modifions les poids des unités influentes et nous utilisons une méthode de calage pour obtenir un ensemble unique de poids d'estimation robustes soit nous modifions les valeurs des unités influentes. À la section 5, nous discutons de l'estimation de l'erreur quadratique moyenne (EQM). À la section 6, nous évaluons certaines propriétés de la méthode proposée au moyen d'une étude en simulation portant sur une population finie asymétrique créée à partir de données d'enquête réelles. Enfin, à la dernière section, nous tirons certaines conclusions.

Soit un vecteur de variables auxiliaires x disponible pour toutes les unités de l'échantillon s et pour lequel on connaît les totaux de population $t_x = \sum_{k \in U} x_k$. Soit, en outre, X la matrice contenant N lignes dont la k^{e} est égale à x_k . Le vecteur x peut ou non contenir certaines variables du vecteur z de variables de plan de sondage. Avant de parler de la robustesse, nous décrivons les deux hypothèses (voir A1 et A2 ci-après) au sujet desquelles la robustesse est souhaitée. Puis, nous expliquons brièvement comment nous les validons.

n'est pas robuste. L'inefficacité relative de l'estimateur GREG comparativement à l'estimateur BLU est due à une forte dispersion des poids de sondage. Le fait que la variabilité des poids de sondage puisse accroître la variance d'un estimateur est bien connu (voir, par exemple, Rao 1966; DuMouchel et Duncan 1983; Pfefferman 1993; Korn et Graubard 1999, chapitre 4; Elliott et Little 2000; Kallott et Flores-Cervantes 2003) et n'est pas rare dans le cas des enquêtes auprès des ménages, à cause des nombreux redressements des poids avant le calage (Kish 1992; Kallott et Flores-Cervantes 2003). Ce problème est souvent traité par troncation des poids de sondage les plus élevés (Potter 1988, 1990, 1993; Stokes 1990).

Pour obtenir un estimateur robuste au plan de sondage quand la variabilité des poids de sondage est forte, nous proposons un compromis entre les estimateurs GREG et BLU qui repose sur la technique des moindres carrés (LS pour *Least Squares*) pondérés. Cet estimateur de compromis présente un biais par rapport au plan de sondage plus faible que celui de l'estimateur BLU quand l'hypothèse de plan ignorable n'est pas satisfaite et, simultanément, est plus efficace que l'estimateur GREG quand cette hypothèse est vérifiée. Il offre aussi une certaine protection contre les écarts par rapport aux hypothèses du modèle. L'échantillonnage équilibré (Royall et Herson 1973) et le calage non paramétrique (Chambers, Dorfman et Wehrly 1993) sont d'autres méthodes qui protègent contre certaines formes d'erreur de spécification du modèle (voir aussi Valliant, Dorfman et Royall 2000, chapitres 3, 4 et 11). Toutefois, aucune de ces méthodes ne protège contre les données aberrantes, qui peuvent être considérées comme la conséquence d'une erreur de spécification du modèle. Dans le cas de l'estimation fondée sur un modèle, l'idée qui sous-tend la technique d'estimation M a été avancée pour élaborer des estimateurs robustes aux données aberrantes pour remplacer l'estimateur BLU (Chambers 1986; Lee 1991; Welsh et Ronchetti 1998). Dans le cas de l'estimation fondée sur le plan de sondage, la technique d'estimation M a aussi été utilisée pour élaborer des estimateurs robustes aux données aberrantes pour remplacer l'estimateur GREG (Gwet et Rivest 1992; Hülliger 1995, 1999; Duchesne 1999; Zaslavsky, Schenker et Belin 2001). L'estimation M est également discutée dans l'article bibliographique de Lee (1995) et une comparaison empirique de plusieurs estimateurs robustes aux données aberrantes figure dans Gwet et Lee (2000).

Les paramètres de population finie sont souvent fort sensibles à la présence de données aberrantes dans la population, contrairement aux paramètres de modèle (population infinie), qui ne dépendent habituellement pas de ces données. Par conséquent, le problème de la robustesse aux données aberrantes n'est pas le même pour les populations

unités de la population, la pratique consiste à sélectionner à partir de la population finie un échantillon aléatoire s de taille n conformément à un plan d'échantillonnage probabiliste $p(s|Z)$. La matrice d'information sur le plan Z contient N lignes dont la k^e est égale à z_k^T , et z est un vecteur de variables auxiliaires disponible à l'étape de l'établissement du plan de sondage. Ceci n'empêche pas de supposer que la population finie proprement dite provient d'un modèle, comme cela est explicitement le cas lorsqu'elle est choisie afin de faire des inférences fondées sur un modèle. Sous ce type d'inférences, Royall (1976) a établi le meilleur estimateur (ou prédicteur) linéaire sans biais (BLU pour *Best Linear Unbiased*) f_B^T de t_y (voir aussi Valliant, Dorfman et Royall 2000, chapitre 2). Cet estimateur est fondé sur les deux hypothèses suivantes : i) le modèle d'estimation sous-jacent à l'estimateur BLU f_B^T est spécifié correctement et ii) le plan de sondage est ignorable en ce qui a trait au modèle d'estimation. Dans ce contexte, un estimateur f_y^T du total de population finie t_y est robuste si sa distribution reste proche de celle de l'estimateur BLU f_B^T quand les deux hypothèses tiennent et qu'il retient de bonnes propriétés quand l'une des hypothèses ou les deux ne sont pas entièrement satisfaites. La robustesse aux écarts par rapport à l'hypothèse (i) est appelée robustesse au modèle, tandis que la robustesse aux écarts par rapport à l'hypothèse (ii) est appelée robustesse au plan de sondage.

Bien que nous considérons des estimateurs robustes construits dans la perspective d'une estimation fondée sur un modèle, nous préférons, dans la mesure du possible, évaluer leurs propriétés par rapport au plan de sondage. Ceci nous permet de choisir les constantes dont dépendent les estimateurs robustes et d'évaluer leur qualité sans devoir nous appuyer sur un modèle et, plus précisément, sans devoir nous appuyer sur un modèle pour les données aberrantes. Cela nous fournit aussi un cadre objectif pour comparer les estimateurs dérivés sous divers modèles. Nous partageons cette préférence pour l'évaluation par rapport au plan de sondage des propriétés des estimateurs fondés sur un modèle avec Little (1983) qui fait remarquer que les propriétés asymptotiques fondées sur un modèle, pourraient être plus utiles pour évaluer les estimateurs que les propriétés asymptotiques fondées sur le plan de sondage

Estimation robuste par la régression généralisée

JEAN-FRANÇOIS BEAUMONT et ASMA ALAYI *

RÉSUMÉ

Le meilleur estimateur (ou prédicteur) linéaire sans biais (BLU) d'un total de population est fondé sur les deux hypothèses suivantes : (i) le modèle d'estimation qui sous-tend l'estimateur BLU est spécifié correctement et (ii) le plan de sondage est ignoré en ce qui concerne le modèle d'estimation. Dans ce contexte, un estimateur est robuste si sa distribution demeure proche de celle de l'estimateur BLU lorsque les deux hypothèses tiennent et s'il retient de bonnes propriétés lorsque l'une des hypothèses ou les deux ne sont pas entièrement satisfaites. La robustesse aux écarts par rapport à l'hypothèse (i) est appelée robustesse au modèle, tandis que la robustesse aux écarts par rapport à l'hypothèse (ii) est appelée robustesse au plan de sondage. On considère souvent que l'estimateur par la régression généralisée (GRBG) est robuste, puisque sa propriété d'être asymptotiquement sans biais par rapport au plan (ASBP) ne dépend ni de l'hypothèse (i) ni de l'hypothèse (ii). Toutefois, si ces deux hypothèses tiennent, l'estimateur GRBG est parfois nettement moins efficace que l'estimateur BLU et, en ce sens, n'est pas robuste. L'inefficacité relative de l'estimateur GRBG comparativement à l'estimateur BLU est due à la grande dispersion des poids de sondage. Afin d'obtenir un estimateur robuste au plan de sondage, nous proposons donc un compromis entre ces deux estimateurs. Cette approche offre aussi une certaine protection contre les écarts par rapport à l'hypothèse (i). Toutefois, elle ne protège pas contre les données aberrantes, qui peuvent être considérées comme la conséquence d'une erreur de spécification du modèle. Pour traiter les données aberrantes, nous utilisons la technique de l'estimation M généralisée pondérée pour réduire l'influence des unités pour lesquelles les résidus pondérés de population sont importants. Nous proposons deux moyens pratiques de mettre en œuvre les estimateurs M dans le cas d'enquêtes polyvalentes; soit nous modifions le poids des unités influentes et adoptons une approche par calage pour obtenir un ensemble unique de poids d'estimation robustes soit nous modifions les valeurs des unités influentes. Nous évaluons certaines propriétés de l'approche proposée au moyen d'une étude en simulation portant sur une population finie asymétrique créée à partir de données d'enquête réelles.

MOTS CLÉS : Robustesse au plan; robustesse au modèle; estimateur M ; données aberrantes; poids rétrécis; meilleur prédicteur linéaire sans biais.

1. INTRODUCTION

En théorie classique, on peut considérer que les données d'échantillon sont tirées aléatoirement d'une population infinie et émettre des hypothèses au sujet de la loi de distribution inconnue de la population infinie. Autrement dit, on postule un modèle et on cherche à en estimer les paramètres. Dans ce contexte, un estimateur $\hat{\theta}$ d'un paramètre du modèle θ est robuste si sa distribution demeure proche de celle de l'estimateur du maximum de vraisemblance de θ quand les hypothèses du modèle tiennent et s'il retient de bonnes propriétés lorsque les hypothèses du modèle ne sont pas entièrement satisfaites. On suppose souvent que la loi de distribution inconnue de la population infinie est la loi normale et, par conséquent, l'estimateur du maximum de vraisemblance se réduit à l'estimateur par les moindres carrés habituel.

L'existence de données aberrantes dans l'échantillon peut être considérée comme la conséquence d'un écart par rapport à une hypothèse du modèle. On pourrait supposer que la majorité de l'échantillon provient du modèle choisi, mais que certaines données, appelées données aberrantes,

proviennent d'un modèle différent. Par conséquent, l'existence de ce genre de données aberrantes dans l'échantillon pourrait introduire un biais et augmenter la variance de l'estimateur par les moindres carrés des paramètres du modèle sélectionné. Les valeurs aberrantes pourraient aussi être le résultat d'une distribution fortement asymétrique. Le cas échéant, l'estimateur par les moindres carrés sans biais, mais peut être très inefficace à cause d'un écart par rapport à l'hypothèse habituelle de normalité. L'existence de données aberrantes dans l'échantillon pourrait aussi être le résultat d'erreurs de mesure. Cependant, nous supposons dans le reste de l'article que les données ont été vérifiées et corrigées, au besoin, et qu'aucune erreur de mesure ne persiste. L'estimation robuste aux données aberrantes pour les populations infinies a été étudiée en détail (pour une revue, voir Huber 1981, ou Hampel, Ronchetti, Rousseeuw et Stahel 1986).

En théorie de l'échantillonnage, on cherche habituellement à estimer des paramètres de population finie, tels que le total, $t_y = \sum_{k \in U} y_k$, d'une variable d'intérêt y pour une population finie U de taille N . Comme il n'est habituellement pas possible d'observer la variable y pour toutes les

- LOHR, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks/Cole.
- PARK, I., et LEE, H. (2001). The design effect: do we know all about it? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, CD-ROM.
- PARK, I., et LEE, H. (2002). A revisit of design effects under unequal probability sampling. *The Survey Statistician*, 46, 23-26.
- SÄRNÄLÄ, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SPENCER, B.D. (2000). Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection. *Techniques d'enquête*, 26, 137-138.
- THOMSEN, I., TESFU, D. et BINDER, D.A. (1986). Estimation of Measures of Size. *International Statistical Review*, 54, 343-349.
- WESTAT (2001). *WestVar 4.0 User's Guide*. Rockville, MD: Westat, Inc.
- YAMANE, T. (1967). *Elementary Sampling Theory*. New Jersey: Prentice-Hall.
- HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. Vol. I, New York: John Wiley & Sons, Inc.
- HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. Vol. II, New York: John Wiley & Sons, Inc.
- JUDKINS, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- KISH, L. (1987). Weighing in Delft. *The Survey Statistician*, Juin 1987.
- KISH, L. (1992). Weighing for unequal p_i . *Journal of Official Statistics*, 8, 183-200.
- KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.
- JANG, D. (2001). On procedures to summarize variances for survey estimates. *Proceedings of the Survey Research Methods of the American Statistical Association*, CD-ROM.
- LEHTONEN, R., et PAHKINEN, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Tableau 1

Comparaison des effets de plan pour le total pondéré et la moyenne pondérée à l'aide d'un sous-ensemble du fichier de données sur les adultes provenant de la troisième National Health and Nutrition Examination Survey (NHANES III) américaine

Caractéristique	Moyenne		Total		$\Delta p(Y)$	$2cv_p(Y)$					
	Estimation	Def ²	Estimation	Def ²							
A fumé 100+ cigarettes au cours de la vie?	Oui	0,53	4,13	0,014	98 397 795	31,31	0,038	0,944	0,20	4,83	-0,58
Fait du diabète?	Oui	0,05	1,75	0,040	9 783 307	1,92	0,042	4,246	-0,34	1,07	-0,31
	Non	0,95	1,75	0,002	176 341 218	393,47	0,033	0,236	0,34	19,35	-5,53
Fait de l'hypertension	Oui	0,23	3,42	0,024	42 939 866	7,96	0,037	1,826	-0,18	2,50	-0,37
	Non	0,77	3,42	0,007	143 184 660	78,44	0,034	0,548	0,18	8,32	-1,22
Race/Groupe ethnique	Afro- américain*	0,12	7,64	0,054	21 567 028	4,21	0,040	2,762	-0,67	1,65	-0,11
	Hispanique*	0,05	6,70	0,079	9 550 326	6,48	0,078	4,300	-0,24	1,06	-0,08
	Masculin	0,48	1,40	0,009	88 725 967	19,18	0,033	1,048	-0,11	4,35	-1,55
	Féminin	0,52	1,40	0,008	97 398 559	25,39	0,034	0,954	0,11	4,77	-1,70
Nombre de cigarettes fumées par jour	-	5,25	6,42	0,037	977 225 826	10,51	0,047	2,044	-0,09	2,23	-0,17
Taille de la population	-	-	-	-	186 124 526	-	0,032	-	-	-	-

Nota : * indique les cas où l'effet de plan est plus faible pour Y que pour $\frac{Y}{\Delta}$

Nota : * indique les cas où l'effet de plan est plus faible pour Y que pour \bar{Y} .

5. CONCLUSION

Nous avons étudié les effets de plan des deux estimateurs les plus répandus de la moyenne et du total de population dans le cas des enquêtes par sondage sous divers plans d'échantillonnage avec remise. À notre avis, l'utilisation d'un échantillonnage avec remise ne constitue pas forcément une limite grave, car elle permet de voir les choses plus clairement, sans embrouiller les calculs à cause des complications probablement inutiles des plans d'échantillonnage sans remise. En outre, l'effet de la correction pour population finie s'annule en grande partie dans notre formule de l'effet de plan, si bien que les résultats sont fort comparables aux effets de plan classiques pour l'échantillonnage sans remise. Par conséquent, nos résultats devraient être utiles en pratique. Nous résumons les plus importants ci-après.

La formule approximative bien connue de l'effet de plan proposée par Kish pour les estimateurs pondérés de la moyenne (type quotient) ne se généralise facilement ni en forme et ni en concepts aux problèmes plus généraux, particulièrement les estimateurs pondérés du total, contrairement à ce que pensent de nombreuses personnes. En fait, \bar{Y} et Y ont souvent des effets de plan fort différents, à moins que le plan d'échantillonnage soit autopondéré ou que les poids d'échantillonnage soient calés sur la taille connue de population. En outre, l'effet de plan n'est généralement pas indépendant de la distribution de la variable étudiée, même pour l'estimateur de la moyenne, sans parler de l'estimateur du total. De surcroît, la corrélation de la variable étudiée et des poids utilisés pour l'estimation peut être un déterminant important de l'effet de plan. Par conséquent, outre ce qu'il

BIBLIOGRAPHIE

- APOSTOL, T.M. (1974). *Mathematical Analysis*. 2^{ème} Ed. Reading, MA: Addison-Wesley.
- BARRON, E.W., et FINCH, R.H. (1978). Design Effects in a complex multistage sample: The Survey of Low Income Aged and Disabled (SLIAD). *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 400-405.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3^{ème} Ed. New York: John Wiley & Sons, Inc.
- CORNFIELD, J. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health*, 41, 654-661.
- GABLER, S., HAEDER, S. et LAHIRI, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Les auteurs remercient Louis Rizza à Westat, un éditeur associé et deux arbitres pour leurs commentaires utiles et leurs suggestions sur une version antérieure de cet article.

REMERCIEMENTS

est destiné à évaluer au départ, l'effet de plan mesure non seulement l'effet d'un plan d'échantillonnage complexe sur une statistique particulière, mais aussi les effets de la distribution de la variable étudiée et de ses relations avec le plan d'échantillonnage sur la statistique. Puisque les projections applicables à des données d'enquête complexes calculent systématiquement l'effet de plan, il semble approprié d'avertir leurs utilisateurs de ces faits assez obscurs au sujet de l'effet de plan.

effets de plan égaux, est la droite de référence. Comme le montre le graphique, la comparaison n'est pas nette. Si $R_p(\hat{Y}, \hat{M}) < 0$, $\text{Def}_p^2(\hat{Y}) \geq \text{Def}_p^2(\hat{Y})$ pour un petit $\text{Def}_p^2(\hat{Y})$, mais la relation est inversée à mesure que $\text{Def}_p^2(\hat{Y})$ s'accroît.

Hansen et coll. (1953, volume I, pages 338–339) indiquent que $R_p(\hat{Y}, \hat{M})$ est souvent proche de 0. Dans ces

conditions, l'expression (4.25) se simplifie en $\text{Def}_p^2(\hat{Y}) \equiv \text{Def}_p^2(\hat{Y}) [1 + \text{CV}_p^2(\hat{M}) / \text{CV}_p^2(\hat{Y})]$, d'où nous obtenons $\text{Def}_p^2(\hat{Y}) \geq \text{Def}_p^2(\hat{Y})$. Ce cas particulier a été étudié par Jang (2001). Cependant, cela ne semble pas nécessaire, comme le montre l'exemple suivant.

Exemple 4.6 Pour illustrer la relation entre les effets de plan pour \hat{Y} et \hat{Y} , nous avons utilisé un ensemble de données sur les adultes provenant de la troisième National Health and Nutrition Examination Survey américaine (NHANES III) qui est fourni comme fichier de démonstration dans la version 4.0 de WesVar. La NHANES III est une enquête avec examen médicale nationale à grande échelle fondée sur un plan d'échantillonnage stratifié à plusieurs degrés, pour lequel la méthode des répliques répétées équilibrées (BRR) modifiée de Fay est employée pour l'estimation de la variance. (Voir Judkins 1990, pour plus de précisions sur la méthode de Fay.) Nous n'avons

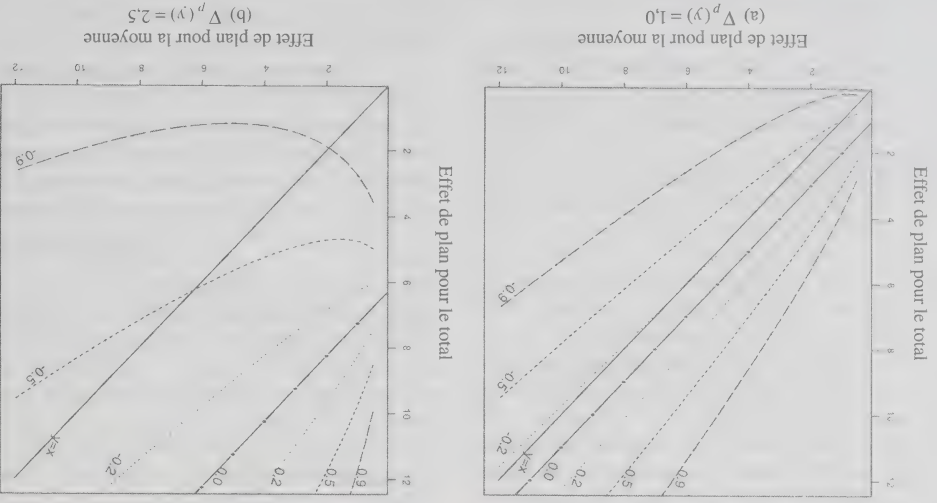


Figure 1. Traces de $\text{Def}_p^2(\hat{Y})$ en fonction de $\text{Def}_p^2(\hat{Y})$ pour a) $V_p(Y) = 1.0$, b) $V_p(Y) = 2.5$. La droite en trait plein correspond à $\text{Def}_p^2(\hat{Y}) = \text{Def}_p^2(\hat{Y})$. Les autres courbes correspondent à $R_p(\hat{Y}, \hat{M}) = -0.9, -0.5, -0.2, 0, 0.2, 0.5$ et 0.9 , respectivement.

Westat (2001).

Le tableau 1 donne les effets de plan pour \hat{Y} et \hat{Y} , et les termes constituant (4.23) pour les caractéristiques choisies. Notons que $V_p(Y)$ diminue de façon monotone en CV_Y . Sachant que $m = 19\,793$ et $\text{cv}_p(\hat{M}) = 3.2\%$. Bien que $V_p(Y)$ ait tendance à être le facteur déterminant de la différence entre les effets de plan, $R_p(\hat{Y}, \hat{M})$ peut jouer un rôle important quand il est négatif. Par exemple, pour deux races/groupe ethniques, à savoir les Afro-américains et les Hispaniques, les valeurs négatives -0.67 et -0.24 de $R_p(\hat{Y}, \hat{M})$ sont les causes de $\text{Def}_p^2(\hat{Y}) < \text{Def}_p^2(\hat{Y})$. Certains effets de plan pour \hat{Y} sont énormes. Il n'en est pas ainsi dans le cas des poids finaux stratifiés à posteriori de la NHANES III, avec lesquels on obtient le même effet de plan pour \hat{Y} et \hat{Y} . Ce résultat souligne l'importance du calage du redressement des poids pour l'estimation des totaux.

voir, écrivons τ sous forme d'une fonction de certaines quantités b_i associées aux UPE comme suit :

$$\tau(b_i) = \frac{(N-1)S_2^{yB} - \sum_{i=1}^N b_i S_2^{yB}}{(N-1)S_2^{yB} + \sum_{i=1}^N (M_i - 1)S_2^{yB}}.$$

Alors, nous obtenons le τ de Thomsen et coll. (1986) avec $b_i = 1$, le τ de l'exemple 4.5 avec $-1/(m_0 - 1)$, et ô de la section 4.2 avec $(M_i - 1)/\{\sum_{i=1}^N (M_i - 1)/(N - 1)\}$. En égalant la formule de Kish (4.18) pour \hat{Y} à (4.19) pour \bar{Y} , ils ont manifestement oublié de tenir compte du fait que les effets de plan pour \bar{Y} et \hat{Y} peuvent être fort différents.

Pour les cas plus généraux, Kish (1987) a établi la formule bien connue suivante pour \hat{Y} :

$$\text{Def}_2^{\text{Kish}}(\hat{Y}) = \frac{\left(\sum_{g=1}^G n_g^2 w_g^2 \right)}{\left[1 + p(\underline{m} - 1) \right]} = (1 + cv_2^w) [1 + p(\underline{m} - 1)].$$

Il l'a obtenue en appliquant (4.5) (ou (4.6)) et (4.18) de façon récursive afin d'intégrer les effets de la mise en grappes ainsi que des poids inégaux. Gähler, Haeder et Lahiri (1999) ont justifié la formule susmentionnée pour \hat{Y} en utilisant un modèle de superpopulation défini pour la classification croisée de N grappes et G classes de pondération. Cependant, on ne peut exposer la différence entre les effets de plan pour \bar{Y} et \hat{Y} selon une approche fondée sur un modèle de ce type, puisque \bar{Y}_g est traitée comme une variable aléatoire tandis que w_g est fixe. Sous cette approche, $\text{Def}_2^{\text{Kish}}(\bar{Y})$ ne diffère de $\text{Def}_2^{\text{Kish}}(\hat{Y})$ que par un facteur $(\bar{M}/M)^2$, alors que la différence réelle peut être nettement plus prononcée, comme nous l'avons montré dans le présent article (par exemple, expressions (3.7) et (4.23)).

4.5 Cas plus généraux

La pondération des données d'enquête nécessite non seulement des poids d'échantillonnage, mais aussi l'application de diverses méthodes de redressement de la pondération, comme la stratification a posteriori, l'ajustement proportionnel itératif (raking) et la correction pour la non-réponse. Nous considérons ces cas généraux ici.

Nous pouvons réécrire l'approximation de premier ordre de Taylor de l'estimateur pondéré de la moyenne $\hat{Y} = \bar{Y}/\bar{M}$ donné par (3.2) sous la forme $(\bar{Y} - Y)/Y \equiv (\hat{Y} - Y)/\bar{Y} + (\bar{M} - M)/\bar{M}$. En prenant la variance des deux membres de l'équation, nous obtenons

$$CV_2^p(\hat{Y}) \equiv CV_2^p(\hat{Y}) + CV_2^p(\bar{M}) + 2R_p(\hat{Y}, \bar{M}) CV_p(\hat{Y}) CV_p(\bar{M}), \quad (4.22)$$

où $CV_2^p(\bar{Y}), CV_2^p(\hat{Y}), CV_2^p(\bar{M})$ et M respectivement et $R_p(\hat{Y}, \bar{M})$ est le coefficient de corrélation de \bar{Y} et \bar{M} par rapport au plan d'échantillonnage complexe p et tout redressement de la pondération. Puisque les variances relatives des simples total et moyenne d'échantillon \bar{Y}_{cas} et \bar{Y}_{cas} sont $CV_2^{\text{cas}}(\bar{Y}_{\text{cas}}) = CV_2^{\text{cas}}(\bar{Y}_{\text{cas}}) = m^{-1} CV_2^{\text{cas}}$ sous casar de taille m , il découle de (4.22) que

$$\text{Def}_2^p(\hat{Y}) \equiv \text{Def}_2^p(\hat{Y}) + 2R_p(\hat{Y}, \bar{M}) \Delta^p(\bar{Y}) \text{Def}_2^p(\bar{Y}) + \Delta^p(\bar{Y})^2, \quad (4.23)$$

où $\Delta^p(\bar{Y}) = CV_p(\bar{M})/CV_p(\bar{Y}_{\text{cas}})$ est non négatif. À titre d'illustration, considérons une variable binaire y , où $CV_2^p \equiv (1 - Y)/\bar{Y}$ et, donc, $\Delta^p(\bar{Y})$ peut être arbitrairement grand quand \bar{Y} s'approche de 1 ou petit quand \bar{Y} s'approche de zéro en supposant que $CV_p(\bar{M}) \neq 0$. Si $\Delta^p(\bar{Y})$ est quasi nul, les deux effets de plan sont presque égaux. Sinon, l'un est plus grand que l'autre dépendant des valeurs de $\Delta^p(\bar{Y})$ et de $R_p(\bar{Y}, \bar{M})$. Si les poids d'échantillonnage sont calés sur la taille connue de population M , \bar{Y} et \bar{Y} ont le même effet de plan, puis que $\bar{M} = M$ et $CV_p(\bar{M}) = 0$. Dans ce cas, le calage n'influe pas sur \bar{Y} , mais $\bar{Y} = M/\bar{Y}$, qui est un estimateur par le quotient. Notons que nous pouvons utiliser les méthodes de stratification a posteriori ou d'ajustement proportionnel itératif (raking) si nous disposons d'information sur la taille de population au niveau de la sous-population et que nous obtenons également des effets de plan équivalents. Néanmoins, en général, nous avons $\text{Def}_2^p(\bar{Y}) \geq \text{Def}_2^p(\hat{Y})$ si

$$R_p(\hat{Y}, \bar{M}) \geq -\frac{2 \text{Def}_2^p(\bar{Y})}{1 - \Delta^p(\bar{Y})} \quad \text{ou} \quad R_p(\hat{Y}, \bar{M}) \geq -\frac{2 CV_p(\bar{Y})}{1 - CV_p(\bar{M})}, \quad (4.24)$$

et inversement.

Il est instructif d'examiner certains cas particuliers. Par exemple, si $R_p(\hat{Y}, \bar{M}) \geq 0$, alors $\text{Def}_2^p(\bar{Y}) > \text{Def}_2^p(\hat{Y})$, mais une corrélation négative (par exemple $R_p(\hat{Y}, \bar{M}) < 0$) ne donne pas nécessairement lieu à $\text{Def}_2^p(\bar{Y}) \leq \text{Def}_2^p(\hat{Y})$. Pour un cas particulier de $R_p(\hat{Y}, \bar{M}) = 0$, la différence est donnée par

$$\text{Def}_2^p(\bar{Y}) - \text{Def}_2^p(\hat{Y}) \equiv \frac{CV_2^p(\bar{Y})}{CV_2^p(\bar{M})} \cdot \frac{CV_2^{\text{cas}}(\bar{Y}_{\text{cas}})}{CV_2^{\text{cas}}(\bar{Y}_{\text{cas}})}. \quad (4.25)$$

La figure 1 illustre la relation entre les deux effets de plan. L'expression (4.23) est représentée graphiquement pour certaines valeurs fixes de $R_p(\hat{Y}, \bar{M})$ et de $\Delta^p(\bar{Y})$. La droite en trait plein passant par l'origine, qui représente des

comme l'échantillonnage en grappes ppi. Nous pouvons appliquer les expressions (4.11) et (4.12) à des probabilités de sélection arbitraires p_i , où les p_i sont fixées de façon à être proportionnelles à une mesure de taille donnée $Z_i \geq 0$. La différence entre les effets de plan pour Y et \bar{Y} est expliquée par Δ_a dans (3.9), ou autrement

$$\Delta_a = \frac{m'}{m'} \sum_{i=1}^s \frac{1}{N} w_i \bar{Q}_i^2 M_2 \left[\left(\frac{\bar{Y}}{Y} \right)^2 - \left(\frac{\bar{D}_i}{D_i} \right)^2 \right]. \quad (4.16)$$

Le terme \bar{Q}_i de (4.16) représente l'effet de p_i sur l'estimation de la variance lorsqu'on utilise une autre mesure de taille que la taille réelle des grappes M_i . Thomsen, Tesfu et Binder (1986) ont considéré, entre autres facteurs, l'effet d'une mesure de taille périinée sous échantillonnage à deux degrés avec échantillonnage aléatoire simple d'éléments à la deuxième étape. Nous y reviendrons à la section 4.4.

4.3 Plans d'échantillonnage autopondérés

Dans le cas d'un échantillonnage autopondéré, chaque élément échantillonné a le même poids, si bien que les estimateurs du total et de la moyenne ont tous deux une forme simple. Ils sont donnés par $\bar{Y} = y/f$ et $\bar{X} = y/m$, où $f = m/M$ est la fraction d'échantillonnage globale et $y = \sum_{i=1}^m y_{ij}$ est le total d'échantillon. Alors, comme dans le cas de l'échantillonnage aléatoire simple, ainsi que le montre (3.4), les deux estimateurs ont le même effet de plan.

Un plan d'échantillonnage autopondéré peut être appliqué de diverses façons par synchronisation des méthodes d'échantillonnage de premier et de deuxième degré (par exemple, Kish 1965, section 7.2). Par exemple, si l'on utilise un échantillonnage avec probabilités égales pour le premier degré, alors l'échantillonnage de deuxième degré devrait se faire selon une méthode à probabilités égales avec une fraction d'échantillonnage uniforme pour toutes les UPE. À titre de cas particulier, où on sélectionne un échantillon aléatoire simple d'UPE de taille égale (c'est-à-dire $M_i = M_0$ pour tout i), Hansen et coll. (1953, volume II, pages 162 – 163) montrent que

$$CV_2^p(\bar{Y}) \equiv \frac{1}{1} CV_2^y [1 + p(m-1)], \quad (4.17)$$

où $CV_2^p(\bar{Y}) = V^p(\bar{Y})/\bar{Y}^2$ est la variance relative de \bar{Y} sous le plan d'échantillonnage p et p est le coefficient de corrélation intragrappe tel que défini dans (4.15c). Puisque la variance relative de \bar{Y} sous casar est $m^{-1} CV_2^y$, la formule approximative bien connue de l'effet de plan pour \bar{Y} sous un plan d'échantillonnage autopondéré s'ensuit immédiatement sous la forme

$$Def_2^p(\bar{Y}) = 1 + p(m-1). \quad (4.18)$$

Pour les plans d'échantillonnage en grappes à un degré, nous avons montré des formes semblables données par (4.15a) et (4.15b) (voir aussi Yamane 1967, section 8.7). Hansen et coll. (1953, volume II, page 204) montrent en outre que $CV_2^p(\bar{Y}) = CV_2^p(\bar{Y})$ pour un plan d'échantillonnage fondé sur l'échantillonnage aléatoire simple aux deux étapes, ce qui implique que Y et \bar{Y} ont le même effet de plan.

4.4 Échantillonnage à deux degrés avec probabilités intégrales

Considérons l'exemple qui suit.

Exemple 4.5 Plan d'échantillonnage à deux degrés où n UPE sont sélectionnées avec remise avec probabilité p_i et un échantillon aléatoire simple de même taille de $m_0 \geq 2$ éléments est sélectionné avec remise à partir de chaque UPE sélectionnée. À l'aide de calculs et de simplifications ordinaires, nous pouvons montrer que

$$Def_2^p(\bar{Y}) \equiv 1 + (m_0 - 1)\tau + W_y^*, \quad (4.19)$$

où

$$\tau = \frac{(N-1)S_{yB}^2 + \sum_{i=1}^N (m_0 - 1)^{-1} S_{y_i}^2}{(N-1)S_{yB}^2 + \sum_{i=1}^N (M_i - 1)^{-1} S_{y_i}^2}, \quad (4.20)$$

$S_{y_i}^2 = (M_i - 1)^{-1} \sum_{j=1}^{m_0} (y_{ij} - \bar{Y}_i)^2$, $W_y^* = W_y/V^{swr}(\bar{Y}) = (m_0/CV_2^y) \sum_{i=1}^N (\bar{Q}_i^2/p_i M_i^2)(\bar{Y}_i^2/Y^2)(1 + CV_2^{y_i}/m_0)$, et $CV_2^{y_i} = S_{y_i}^2/\bar{Y}_i^2$ de la variable y_i . De la même façon,

$$Def_2^p(\bar{X}) \equiv 1 + (m_0 - 1)\tau + W_x^*, \quad (4.21)$$

où $W_x^* = W_x/V^{casar}(\bar{X}) = (m_0/CV_2^x) \sum_{i=1}^N (\bar{Q}_i^2/p_i M_i^2)(\bar{D}_i^2/X^2)(1 + CV_2^{x_i}/m_0)$, et $\bar{D}_i = (D_i/Y)^2$ de la variable transformée $d(d_i = y_{ij} - X)$ de façon analogue à \bar{Y}_i et $CV_2^{x_i}$, respectivement. (Les calculs détaillés de l'établissement des expressions (4.19) et (4.21) sont donnés par (4.19) et (4.21) se réduit à (3.7) ou (4.16). L'échantillonnage de deuxième degré ne contribue pas à la différence.

Si nous revenons à Thomsen et coll. (1986), qui ont étudié l'effet de l'utilisation d'une mesure périinée de taille sur la variance, nous voyons que la discussion qui précède concernant Y est équivalente à la leur. La seule différence tient au fait qu'ils émettent l'hypothèse d'un plan d'échantillonnage sans remise à la deuxième étape. Notons, toutefois, que la définition de τ dans Thomsen et coll. (1986) diffère légèrement de (4.20) et de δ à la section 4.2. Cependant, ces définitions sont étroitement liées. Pour le

Nous observons que l'effet de plan pour \bar{Y} diffère de celui pour \bar{Y} dans le deuxième terme contenant $D_i^p = \sum_{j=1}^N (Y_j^p - \bar{Y})$ au lieu de \bar{Y}_i . En outre, notons que la quantité $\delta = \delta^p(\bar{Y})$ est le coefficient de détermination ajusté (R_{adj}^2) dans le contexte de l'analyse par régression. On peut le considérer comme une mesure d'homogénéité. Pour une discussion plus approfondie de δ , voir Sâmdal et coll. (1992, pages 130-131) et Lohr (1999, page 140).

Exemple 4.3 Échantillonnage aléatoire simple à un degré de grappes. Dans cet exemple, si $p_i = 1/N$ pour tout $i = 1, \dots, N$, les deux effets de plan de sondage donnés par (4.11) et (4.12) se réduisent, respectivement, à

$$\text{Def}_2^p(\bar{Y}) = \left(\frac{N}{N-1} \right) \left(1 + \frac{N-1}{M} \delta \right) + \frac{1}{N} \sum_{i=1}^N (M_i - \bar{M}) \left(\frac{\bar{M}_i}{\bar{Y}} \right) \left(\frac{\bar{Y}}{\bar{D}_i} \right) \quad (4.13)$$

$$\text{Def}_2^p(\bar{Y}) \equiv \left(\frac{N}{N-1} \right) \left(1 + \frac{N-1}{M} \delta \right) + \frac{1}{N} \sum_{i=1}^N (M_i - \bar{M}) \left(\frac{\bar{M}_i}{\bar{D}_i} \right) \left(\frac{\bar{Y}}{\bar{D}_i} \right), \quad (4.14)$$

où $\bar{M} = M/N$. Puisque $\text{Def}_2^p(\bar{Y}) - \text{Def}_2^p(\bar{Y}) \propto \sum_{i=1}^N (M_i - \bar{M}) (2\bar{Y}_i - \bar{Y})$, l'inégalité entre les effets de plan pour \bar{Y} et \bar{Y} dépend de la loi conjointe de \bar{Y}_i et M_i .

Exemple 4.4 Échantillonnage aléatoire simple à un degré de grappes de même taille. Dans ce cas-ci, nous avons $M_i \equiv M_0$ et $p_i = 1/N$ pour tout $i = 1, \dots, N$ et nous pouvons approximer les deux effets de plan donnés par (4.13) et (4.14) par la même quantité donnée par

$$(4.15a) \quad \left(\frac{N}{N-1} \right) \left[1 + \frac{N}{N(M_0-1)} \delta \right],$$

Pour tenir compte de l'effet de la mise en grappes sur puisque $M_i - \bar{M} = 0$ pour tout $i = 1, \dots, N$.

L'estimation de la variance, on utilise souvent la forme la plus simple d'échantillonnage aléatoire simple en grappes à un degré comme dans l'exemple 4.4. Consulter, par exemple, Cochran (1977, section 9.4), Lehtonen et Pakkinen (1995, page 91) et Lohr (1999, section 5.2.2). Bien que ces auteurs aient adopté un plan d'échantillonnage sans remise, par souci de simplicité et de cohérence, nous comparons leurs formules à notre formule hypothèse d'échantillonnage avec remise. De surcroît, la comparaison est valide parce que leurs formules intègrent la correction pour population finie au numérateur ainsi qu'au dénominateur, si

section 9.4) obtient la formule

$$\text{Def}_2^p(\bar{Y}) = \frac{NM_0 - 1}{M_0(N-1)} [1 + (M_0 - 1)p], \quad (4.15b)$$

où p est le coefficient de corrélation intragrappe défini par

$$p = \frac{2 \sum_{i=1}^N \sum_{j>k=1}^N (Y_j^p - \bar{Y})(Y_k^p - \bar{Y})}{(M_0 - 1) \sum_{i=1}^N \sum_{j=1}^N (Y_j^p - \bar{Y})^2}, \quad (4.15c)$$

En réécrivant $\sum_{i=1}^N [\sum_{j=1}^N (Y_j^p - \bar{Y})^2] = M_0(N-1)S_2^{y^p}$ et $\sum_{i=1}^N \sum_{j>k=1}^N (Y_j^p - \bar{Y})(Y_k^p - \bar{Y}) = (N-1)S_2^{y^p} + N(M_0 - 1)S_2^{y^p}$, il est facile de montrer que

$$2 \sum_{i=1}^N \sum_{j>k=1}^N (Y_j^p - \bar{Y})(Y_k^p - \bar{Y}) = \sum_{i=1}^N \sum_{j=1}^N (Y_j^p - \bar{Y})^2 - \sum_{i=1}^N \sum_{j=1}^N (Y_j^p - \bar{Y})^2 = (M_0 - 1)S_2^{y^p} - NM_0S_2^{y^p}$$

et, donc, partant de (4.15c), que $p = 1 - (NM_0/NM_0 - 1)\{ (S_2^{y^p}/S_2^y) \equiv \delta$ en supposant que $M_i \equiv M_0$ pour tout $i = 1, \dots, N$, $NM_0/(NM_0 - 1) \equiv 1$. Par conséquent, en supposant de surcroît que $(N-1)/N \equiv 1$ et $(NM_0 - 1)M_0^{-1}(N-1)^{-1} \equiv 1$, les deux formules de l'effet de plan (4.15a) et (4.15b) sont approximativement équivalentes à $1 + (M_0 - 1)\delta$. D'autres auteurs arrivent à la même formule approximative. Il en est ainsi parce que δ et p mesurent essentiellement la même chose, c'est-à-dire l'homogénéité de la grappe. Dans ces conditions, deux estimateurs \bar{Y} et \bar{Y} ont le même effet de plan, tel que discuté à l'exemple 3.2. Notons qu'il s'agit d'un cas simple de plan d'échantillonnage autopondéré.

Sâmdal et coll. (1992, section 8.7) comparent les effets de plan pour deux estimateurs dans les conditions de l'exemple 4.3. Ils établissent aussi une expression simplifiée $1 + (\bar{M} - 1)\delta$ pour (4.13) et (4.14), en supposant qu'on peut ne pas tenir compte des covariances de M_i avec M_i/Y_i^p et M_i/D_i^p . Leur discussion de la différence entre les estimateurs du total et de la moyenne se résume à Δ_a dans l'exemple 3.2. Ils notent aussi que l'effet de plan peut être beaucoup plus important pour le total que pour la moyenne de population, parce que la perte due à l'échantillonnage de grappes est plus importante quand on estime le total que quand on estime la moyenne.

Une pratique courante lorsque les grappes sont de taille inégale consiste à utiliser une méthode d'échantillonnage plus efficace qui tient compte de la différence de taille, mais

où $CV_y^2 = n^{-1} \sum_{i=1}^N (w_i - \bar{w})^2 / \bar{w}^2$ est la variance d'échantillon relative et \bar{w} est la moyenne d'échantillon de w_i . Nous que (4.6) est une approximation sur échantillon de (4.3). Dans le cas d'un plan d'échantillonnage inefficace pour l'estimation de X , l'inefficacité diminue avec l'estimation par le quotient. Considérons maintenant le cas opposé d'une corrélation de la variable y et de la probabilité de sélection p_i , où l'efficacité de Y augmente.

Exemple 4.2 Échantillonnage d'éléments avec probabilités inégales où y_i est corrélée à p_i . Supposons que y_i est reliée linéairement à p_i par $y_i = A + Bp_i + e_i$, où A et B sont les coefficients de régression par les moindres carrés du modèle pour la population (finie) et e_i est le résidu correspondant. En outre, supposons que le modèle de régression soit bien ajusté aux données de population et que la variance de l'erreur soit à peu près homogène, de sorte que $R^{ew} \approx 0$ et $R^{ew} \approx 0$, où R^{ew} et R^{ew} représentent les corrélations de population des paires (e_i, w_i) et (e_i^2, w_i) , respectivement. Par exemple, $R^{ew} = \sum_{i=1}^N (e_i - \bar{e})(w_i - \bar{w}) / ((N-1)S_e S_w)$, où $\bar{e} = \sum_{i=1}^N e_i / N$, S_e et S_w sont les écarts-types de population de e_i et w_i , respectivement. Alors, les effets de plan donné par (4.1) et (4.2) se réduisent à

$$\begin{aligned} \text{Def}_p^2(Y) &\equiv (n\bar{w}/N)(1 - R_y^2) + (n\bar{w}/N - 1) \left(\frac{CV_y^p}{CV_y} - 1 \right) \\ \text{Def}_p^2(\hat{Y}) &\equiv (n\bar{w}/N)(1 - R_y^2) + (n\bar{w}/N - 1) \left(R_y^{\frac{p}{2}} \frac{CV_y^p}{CV_y} - 1 \right) \end{aligned} \quad (4.7)$$

et

$$2R_y^p \leq CV_y^p / CV_y, \quad (4.9)$$

où l'égalité tient si, et uniquement si, $2R_y^p = CV_y^p / CV_y$. De surcroît, l'inégalité est inversée si l'inégalité (4.9)

La condition (4.9) indique que Y a tendance à être moins efficace que \hat{Y} en ce qui concerne la précision quand R_y^p est petit. Donc, nous voyons que R_y^p est un déterminant important de l'effet de plan d'échantillonnage avec probabilités inégales sur Y et \hat{Y} et leur efficacité relative.

4.2 Échantillonnage en grappes à un degré

Considérons un échantillonnage en grappes à un degré, où chaque élément compris dans une grappe échantillonnée est inclus dans l'échantillon, c'est-à-dire $m_i = M_i$ pour tout $i \in s_a$. Étant donné l'absence de la variation d'échantillonnage de deuxième degré, la variance de Y correspond uniquement au premier terme de l'expression (3.1) et peut être décomposée comme suit

$$\sum_{i=1}^N w_i (X_i - p_i X)^2 = \frac{n}{M(N-1)} S_y^2 + \sum_{i=1}^N w_i \bar{Q}_i^2 \bar{Y}^2, \quad (4.10)$$

où $S_y^2 = (N-1)^{-1} \sum_{i=1}^N M_i (Y_i - \bar{Y})^2$ et $\bar{Q}_i^2 = M_i(M_i - p_i M)^2$ pour $i = 1, \dots, N$. Notons que $\bar{Q}_i = 0$ si $p_i = M_i / M$, c'est-à-dire que p_i est proportionnel à la taille de grappe M_i . Notons aussi que S_y^2 est la moyenne quadratique des écarts entre grappes dans une analyse de variance. En représentant la moyenne quadratique des écarts à l'intérieur des grappes par $S_y^2 = (M - N)^{-1} \sum_{i=1}^N \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2$, écrivons $S_y^2 = S_y^2 [1 + \delta(M - N)/(N - 1)]$ avec $\delta = 1 - S_y^w / S_y^2$. Puisque la taille d'échantillon prévue est $m' = nM$, l'effet de plan pour Y peut s'écrire, en partant de (4.10), sous la forme

$$\text{Def}_p^2(Y) = \left(\frac{N-1}{N-1} \right) \left(1 + \frac{M-N}{N} \delta \right) + \frac{nM}{N} \sum_{i=1}^N \frac{CV_y^2}{w_i \bar{Q}_i} \left(\frac{Y}{\bar{Y}} \right)^2. \quad (4.11)$$

De la même façon, l'effet de plan pour \hat{Y} peut être exprimé par

$$\text{Def}_p^2(\hat{Y}) \equiv \left(\frac{N-1}{N-1} \right) \left(1 + \frac{M-N}{N} \delta \right) + \frac{nM}{N} \sum_{i=1}^N \frac{CV_y^2}{w_i \bar{Q}_i} \left(\frac{Y}{D} \right)^2. \quad (4.12)$$

Nous observons que l'effet de plan pour \hat{Y} diffère de celui pour \hat{Y} dans le deuxième terme contenant $D_i^j = \sum_{i=1}^N (Y_i^j - \bar{Y})$ au lieu de \hat{Y}_i . En outre, notons que la quantité $\delta = \delta_p(Y)$ est le coefficient de détermination ajusté (R_{adj}^2) dans le contexte de l'analyse par régression. On peut le considérer comme une mesure d'homogénéité. Pour une discussion plus approfondie de δ , voir Särndal et coll. (1992, pages 130-131) et Lohr (1999, page 140).

Exemple 4.3 Échantillonnage aléatoire simple à un degré de grappes. Dans cet exemple, si $p_i = 1/N$ pour tout $i = 1, \dots, N$, les deux effets de plan de sondage donnés par (4.11) et (4.12) se réduisent, respectivement, à

$$\text{Def}_2^p(Y) = \left(1 + \frac{N}{M - N}\delta\right) \left(\frac{N}{N-1}\right) \left(1 + \frac{N}{M - N}\delta\right)$$

$$+ \frac{1}{N \cdot \text{CV}_Y^2} \sum_{i=1}^N (M_i - \bar{M}) \left(\frac{M_i}{\bar{Y}}\right) \left(\frac{\bar{Y}}{Y_i}\right) \quad (4.13)$$

$$\text{Def}_2^p(Y) \equiv \left(\frac{N}{N-1}\right) \left(1 + \frac{N}{M - N}\delta\right)$$

$$+ \frac{1}{N \cdot \text{CV}_Y^2} \sum_{i=1}^N (M_i - \bar{M}) \left(\frac{M_i}{\bar{Y}}\right) \left(\frac{\bar{Y}}{Y_i}\right), \quad (4.14)$$

où $\bar{M} = M/N$. Puisque $\text{Def}_2^p(Y) - \text{Def}_2^p(\hat{Y}) \rightarrow \sum_{i=1}^N M_i (M_i - \bar{M}) / (2\bar{Y} - \bar{Y})$, l'inégalité entre les effets de plan pour \hat{Y} et \hat{Y} dépend de la loi conjointe de \hat{Y}_i et M_i .

Exemple 4.4 Échantillonnage aléatoire simple à un degré de grappes de même taille. Dans ce cas-ci, nous avons $M_i \equiv M_0$ et $p_i = 1/N$ pour tout $i = 1, \dots, N$ et nous pouvons approximer les deux effets de plan donnés par (4.13) et (4.14) par la même quantité donnée par

$$\left(\frac{N}{N-1}\right) \left(1 + \frac{N}{N(M_0 - 1)}\delta\right), \quad (4.15a)$$

puisque $M_i - \bar{M} = 0$ pour tout $i = 1, \dots, N$.

L'estimation de la variance, on utilise souvent la forme la plus simple d'échantillonnage aléatoire simple en grappes à un degré comme dans l'exemple 4.4. Consulter, par exemple, Cochran (1977, section 9.4), Lehtonen et Pakkinnen (1995, page 91) et Lohr (1999, section 5.2.2). Bien que ces auteurs aient adopté un plan d'échantillonnage sans remise, par souci de simplicité et de cohérence, nous comparons leurs formules à notre formule avec hypothèse d'échantillonnage avec remise. De surcroît, la comparaison est valide parce que leurs formules intègrent la correction pour population finie au numérateur ainsi qu'au dénominateur, si

où p est le coefficient de corrélation intragrappe défini par

$$\text{Def}_2^p(\hat{Y}) = \frac{NM_0 - 1}{M_0(N - 1)} [1 + (M_0 - 1)p] \quad (4.15b)$$

bien que cet effet s'annule essentiellement. Cochran (1977, section 9.4) obtient la formule

$$p = \frac{2 \sum_{i=1}^N \sum_{j>k=1}^N (Y_i^j - \bar{Y})(Y_k^j - \bar{Y})}{(M_0 - 1) \sum_{i=1}^N \sum_{j=1}^N (Y_i^j - \bar{Y})^2}. \quad (4.15c)$$

En réécrivant $\sum_{i=1}^N \sum_{j=1}^N (Y_i^j - \bar{Y})^2 = (NM_0 - 1)S_y^2 + N(M_0 - 1)S_{yw}^2$, il est facile de montrer que

$$2 \sum_{i=1}^N \sum_{j>k=1}^N (Y_i^j - \bar{Y})(Y_k^j - \bar{Y}) = \sum_{i=1}^N \sum_{j=1}^N (Y_i^j - \bar{Y})^2 - \sum_{i=1}^N \sum_{j=1}^N (Y_i^j - \bar{Y})^2 = (M_0 - 1)[(NM_0 - 1)S_y^2 - NM_0 S_{yw}^2]$$

et, donc, partant de (4.15c), que $p = 1 - (NM_0 - 1) / (NM_0 - 1)S_y^2 \equiv \delta$ en supposant que $M_i \equiv M_0$ pour tout $i = 1, \dots, N$, $NM_0 / (NM_0 - 1) \equiv 1$. Par conséquent, en supposant de surcroît que $(N - 1) / N \equiv 1$ et $(NM_0 - 1)M_0^{-1}(N - 1)^{-1} \equiv 1$, les deux formules de l'effet de plan (4.15a) et (4.15b) sont approximativement équivalentes à $1 + (M_0 - 1)\delta$. D'autres auteurs arrivent à la même formule approchée. Il en est ainsi parce que δ et p mesurent essentiellement la même chose, c'est-à-dire l'homogénéité de la grappe. Dans ces conditions, deux estimateurs \hat{Y} et \hat{Y} ont le même effet de plan, tel que discuté à l'exemple 3.2. Notons qu'il s'agit d'un cas simple de plan d'échantillonnage autopondéré.

Särndal et coll. (1992, section 8.7) comparent les effets de plan pour deux estimateurs dans les conditions de l'exemple 4.3. Ils établissent aussi une expression simplifiée $1 + (M - 1)\delta$ pour (4.13) et (4.14), en supposant qu'on peut pas tenir compte des covariances de M_i avec $M_i \bar{Y}^2$ et $M_i \bar{D}_i^2$. Leur discussion de la différence entre les estimateurs du total et de la moyenne se résume à Δ_a dans l'exemple 3.2. Ils notent aussi que l'effet de plan peut être beaucoup plus important pour le total que pour la moyenne de population, parce que la perte due à l'échantillonnage de grappes est plus importante quand on estime le total que quand on estime la moyenne.

Une pratique courante lorsque les grappes sont de taille inégale consiste à utiliser une méthode d'échantillonnage plus efficace qui tient compte de la différence de taille,

où $CV_2^w = n^{-1} \sum_{i=1}^N (w_i - \bar{w})^2 / \bar{w}^2$ est la variance d'échantillon relative et \bar{w} est la moyenne d'échantillon de w_i . Notons que (4.6) est une approximation sur échantillon de (4.3). Dans le cas d'un plan d'échantillonnage inefficace pour l'estimation de X , l'inefficacité diminue avec l'estimation par le quotient. Considérons maintenant le cas opposé d'une corrélation de la variable y et de la probabilité de sélection p_i , où l'efficacité de X augmente.

Exemple 4.2 Échantillonnage d'éléments avec probabilités inégales où y_i est corrélée à p_i . Supposons que y_i est reliée linéairement à p_i par $y_i = A + Bp_i + e_i$, où A et B sont les coefficients de régression par les moindres carrés du modèle pour la population (finie) et e_i est le résidu correspondant. En outre, supposons que le modèle de régression soit bien ajusté aux données de population et que la variance de l'erreur soit à peu près homogène, de sorte que $R^{yw} \equiv 0$ et $R^{pw} \equiv 0$, où R^{yw} et R^{pw} représentent les corrélations de population des paires (e_i, w_i) et (e_i^2, w_i) , respectivement. Par exemple, $R_{vw}^p = \sum_{i=1}^N (e_i - \bar{E})(w_i - \bar{W}) / ((N-1)S_e^p S_w^p)$, où $\bar{E} = \sum_{i=1}^N e_i / N$, S_e^p et S_w^p sont les écarts-types de population de e_i et w_i , respectivement. Alors, les effets de plan donné par (4.1) et (4.2) se réduisent à

$$\begin{aligned} \text{Def}_2^p(\hat{Y}) &\equiv (n\bar{w}/N)(1 - R_2^{yp}) \\ &+ (n\bar{w}/N - 1) \left(\frac{CV_p}{R^{yp}} - \frac{1}{CV_y} \right), \end{aligned} \quad (4.7)$$

$$\text{Def}_2^p(\hat{Y}) \equiv (n\bar{w}/N)(1 - R_2^{yp}) + (n\bar{w}/N - 1) \left(\frac{CV_p}{R^{yp}} - \frac{1}{CV_y} \right), \quad (4.8)$$

et

respectivement, où R^{yp} est la corrélation de variation de y_i et p_i et CV_p est le coefficient de variation de population de p_i (voir Park et Lee (2001), pour la preuve). Il découle de (4.7) et (4.8) que $\text{Def}_2^p(\hat{Y}) \geq \text{Def}_2^p(\hat{Y})$, si, et uniquement si

$$2R^{yp} \leq CV_p / CV_y, \quad (4.9)$$

où l'égalité tient si, et uniquement si, $2R^{yp} = CV_p / CV_y$. De surcroît, l'inégalité est inversée si l'inégalité (4.9) devient opposée.

La condition (4.9) indique que X a tendance à être moins efficace que \hat{Y} en ce qui concerne la précision quand R^{yp} est petit. Donc, nous voyons que R^{yp} est un déterminant important de l'effet de plan d'échantillonnage avec probabilités inégales sur \hat{Y} et \hat{X} et leur efficacité relative.

4.2 Échantillonnage en grappes à un degré

En s'efforçant d'élaborer une expression approximative de l'effet de plan quand y_i est corrélée à p_i , Spencer (2000) a proposé une formule approximative d'échantillon pour X et l'a comparée à la formule approximative de Kish pour le cas particulier où $R^{yp} = 0$. Comme le montre l'exemple 4.2, les deux effets de plan (4.7) et (4.8) ne sont pas égaux à moins que $\bar{W} = N/n$ (voir Park et Lee (2001) pour une discussion plus approfondie et certains exemples numériques). En outre, ce cas particulier donne la même condition que pour l'exemple 4.1 et, par conséquent, les deux formules approximatives de l'effet de plan (4.7) et (4.8) sont équivalentes à (4.4) et (4.3), respectivement.

Considérons un échantillonnage en grappes à un degré, où chaque élément compris dans une grappe échantillonnée est inclus dans l'échantillon, c'est-à-dire $m_i = M_i$ pour tout $i \in s$. Étant donné l'absence de la variation d'échantillonnage de deuxième degré, la variance de \hat{Y} correspond uniquement au premier terme de l'expression (3.1) et peut être décomposée comme suit

$$\sum_{i=1}^N w_i (Y_i - p_i X)^2 = \frac{n}{M(N-1)} S_{yb}^2 + \sum_{i=1}^N w_i \bar{Q}_i^2 \bar{X}^2, \quad (4.10)$$

où $S_{yb}^2 = (N-1)^{-1} \sum_{i=1}^N M_i (Y_i - \bar{Y})^2$ et $\bar{Q}_i^2 = M_i(M_i - p_i M)^{-1}$ pour $i = 1, \dots, N$. Notons que $\bar{Q}_i^2 = 0$ si $p_i = M_i/M$, c'est-à-dire que p_i est proportionnel à la taille de grappe M_i . Notons aussi que S_{yb}^2 est la moyenne quadratique des écarts entre grappes dans une analyse de variance. En représentant la moyenne quadratique des écarts à l'intérieur des grappes par $S_{yw}^2 = (M-N)^{-1} \sum_{i=1}^N \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2$, écrivons $S_{yb}^2 = S_y^2 \{1 + \delta(M-N)/(N-1)\}$ avec $\delta = 1 - S_{yw}^2/S_y^2$. Puisque la taille d'échantillon prévue est $m' = n\bar{M}$, l'effet de plan pour \hat{Y} peut s'écrire, en partant de (4.10), sous la forme

$$\text{Def}_2^p(\hat{Y}) = \left(\frac{N}{N-1} \right) \left(1 + \frac{M-N}{N} \delta \right) + \frac{n\bar{M}}{CV_y^2} \sum_{i=1}^N \frac{w_i \bar{Q}_i^2}{M_i} \left(\frac{Y_i}{\bar{Y}} \right)^2. \quad (4.11)$$

De la même façon, l'effet de plan pour \hat{X} peut être exprimé par

$$\text{Def}_2^p(\hat{X}) \equiv \left(\frac{N}{N-1} \right) \left(1 + \frac{M-N}{N} \delta \right) + \frac{n\bar{M}}{CV_x^2} \sum_{i=1}^N \frac{w_i \bar{Q}_i^2}{M_i} \left(\frac{Y_i}{D} \right)^2. \quad (4.12)$$

aléatoire simple des grappes ne produit aucune différence entre les effets de plan.

À la section 4, nous utilisons les résultats obtenus à la présente section pour discuter d'autres exemples utilisés dans la littérature sur l'échantillonnage.

4. EXEMPLES DE L'EFFET DE PLAN DANS LA LITTÉRATURE SUR L'ÉCHANTILLONNAGE

4.1 Échantillonnage d'éléments avec probabilités inégales

Considérons l'échantillonnage d'éléments avec probabilités inégales sans mise en grappes. La discussion de la section 3 s'applique à cet exemple avec $M_i \equiv 1$ pour tout $i = 1, \dots, N$ et, donc, $m = n$. Par souci de concision, nous utilisons y_i pour représenter la valeur de la variable y et nous supposons que N est grand de sorte que $N/(N-1) \approx 1$. En l'absence de la variation d'échantillonnage de deuxième degré, les effets de plan pour \bar{Y} et \bar{Y} donnés par les expressions (3.5) et (3.6) se réduisent à

$$\text{Def}_2^p(\bar{Y}) \equiv \frac{\sum_{i=1}^N p_i^{-1}(y_i - \bar{Y})^2}{\sum_{i=1}^N N(y_i - \bar{Y})^2} \quad (4.1)$$

et

$$\text{Def}_2^p(\hat{\bar{Y}}) \equiv \frac{\sum_{i=1}^N p_i^{-1}(y_i - \bar{Y})^2}{\sum_{i=1}^N N(y_i - \bar{Y})^2}. \quad (4.2)$$

En outre, considérons un exemple où la variable étudiée y n'est pas corrélée à la probabilité de sélection p_i .

Exemple 4.1 Échantillonnage d'éléments avec probabilités inégales sans corrélation entre y_i et p_i . Si y_i et p_i ne sont pas corrélées, nous pouvons approximer $\sum_{i=1}^N p_i^{-1}(y_i - \bar{Y})^2$ par $N \sum_{i=1}^N (y_i - \bar{Y})^2$, où $\bar{W} = N^{-1} \sum_{i=1}^N w_i^2$. Notons que $E^p(n^{-1} \sum_{i=1}^n w_i^2) = N/n$, $E^p(n^{-1} \sum_{i=1}^n w_i^2) = NW/n$ et $E^p(n^{-1} \sum_{i=1}^n w_i^2) / E^p(n^{-1} \sum_{i=1}^n w_i^2) = \bar{W}/N$. Donc,

$$\text{Def}_2^p(\hat{\bar{Y}}) \equiv n \bar{W} / N$$

$$= E^p \left(n^{-1} \sum_{i=1}^n w_i^2 \right) / E^p \left(n^{-1} \sum_{i=1}^n w_i^2 \right). \quad (4.3)$$

Il est facile de montrer que $n \bar{W} / N \geq 1$ en utilisant l'inégalité de Cauchy-Schwarz (Apostol 1974, page 14). En outre, des calculs ordinaires montrent, en partant de (4.1) et (4.2), que

$$\text{Def}_2^p(\hat{\bar{Y}}) = n \sum_{i=1}^n w_i^2 / \left(\sum_{i=1}^n w_i^2 \right) \left(\sum_{i=1}^n w_i^2 \right)^{-1} = 1 + c v_w^2, \quad (4.6)$$

approximative bien connue donnée par

Plus tard, en posant que, dans (4.5), $n_g = 1$ pour tout g et, donc, $n = G$, Kish (1992) a proposé une formule d'échantillonnage stratifié proportionné.

Un échantillonnage stratifié proportionné à celle sous inégale peut être approximée par le rapport de la variance de \bar{Y} due à une pondération aléatoire de précision n . La logique du calcul susmentionné est que la perte comparativement à la variance hypothétique sous casar de pour mesurer l'augmentation de la variance de \bar{Y}

$$\text{Def}_2^p(\hat{\bar{Y}}) = n \sum_{g=1}^G w_g^2 / \left(\sum_{g=1}^G n_g w_g^2 \right) \left(\sum_{g=1}^G n_g w_g^2 \right)^{-1}. \quad (4.5)$$

Une quantité donnée par

c'est-à-dire $S_{y_g}^2 = S_y^2$ pour tout $g = 1, \dots, G$, il a proposé pondération sont égales à la variance unitaire de y , $n = \sum_{g=1}^G n_g$, et que (2) les G variances de classe de unités d'échantillonnage dans la classe g et que pondération telles que le même poids w_g soit attribué à n_g de taille n tiré avec remise est subdivisé en G classes de non-réponse. En supposant que (1) un échantillon aléatoire de base de sondage ou des redressements pour la inégaux ont une origine « aléatoire », comme des problèmes (1965, 11.7). Celui-ci a considéré les cas où les poids sur l'efficacité de plan donnée par (2.2) est due à Kish

La quantification habituelle de l'effet des poids inégaux y_i et p_i est faible et que $\text{Def}_2^p(\bar{Y}) \geq 1$.

L'exemple 4.1 montre que l'effet de plan a tendance à être plus grand pour \bar{Y} que pour $\hat{\bar{Y}}$ si la corrélation entre y_i et p_i est faible et que $\text{Def}_2^p(\bar{Y}) < 1$.

D'après (4.4), il est clair que $\text{Def}_2^p(\bar{Y}) \geq \text{Def}_2^p(\hat{\bar{Y}})$ si $\text{Def}_2^p(\bar{Y}) \geq 1$ et que l'égalité tient si $\text{Def}_2^p(\bar{Y}) = 1$ ou $\bar{W} = N/n$. En outre, $\text{Def}_2^p(\bar{Y}) < \text{Def}_2^p(\hat{\bar{Y}})$ si

$$\text{Def}_2^p(\bar{Y}) \equiv (1 + c v_y^2) \text{Def}_2^p(\hat{\bar{Y}}) - c v_y^2. \quad (4.4)$$

ou

$$\text{Def}_2^p(\bar{Y}) - \text{Def}_2^p(\hat{\bar{Y}}) \equiv c v_y^2 \left\{ \text{Def}_2^p(\bar{Y}) - 1 \right\}$$

sont pas corrélées. Par conséquent,

et $\sum_{i=1}^N p_i^{-1}(y_i - \bar{Y})(p_i - \bar{p}) \equiv 0$ parce que y_i et p_i ne expression à partir de $\sum_{i=1}^N p_i^{-1}(p_i - \bar{p})^2 = n \bar{W} / N - 1$ où $\bar{p} = N^{-1} \sum_{i=1}^N p_i = 1/N$. Nous obtenons la dernière

$$\text{Def}_2^p(\hat{\bar{Y}}) - \text{Def}_2^p(\bar{Y}) \equiv c v_y^2 \left\{ \sum_{i=1}^N p_i^{-1}(p_i - \bar{p})^2 - 2 \sum_{i=1}^N p_i^{-1}(y_i - \bar{Y})(p_i - \bar{p}) \right\} = c v_y^2 (n \bar{W} / N - 1),$$

ou $CV_z^y = S_z^y / Y^z$ représente la variance de population

$$\text{Def}_2^p(Y) - \text{Def}_2^p(\hat{Y}) \equiv \Delta_a + \Delta_b, \quad (3.7)$$

$$\cdot \left\{ \left[\left(\frac{\lambda}{\hat{D}_l} \right)^q V - \left(\frac{\lambda}{\hat{X}_l} \right)^q V \right] ' W \sum_{N=1}^l \frac{C V_2}{m'} = \Delta^q \right.$$

disparaît si tous les éléments compris dans les grappes sélectionnées sont observés, puis que le plan devient un plan à un seul degré de liberté, on ne peut pas échantillonner aléatoirement une sélection à deux degrés. Il en est ainsi parce que les deux variances $V_p(Y_i)$ et $V_p(Y_j)$ sont équivalentes sous les conditions susmentionnées, c'est-à-dire $(1) V_p(Y_i) = V_p(Y_j) = 0$ si $w_i^{jit} = 1$ pour tous i et j , et $(2) V_p(Y_i) = V_p(Y_j) = M_i^{jit}/m_i$ pour tous i et j . Autrement dit,

$$(6.9) \quad \left. \begin{array}{lcl} {}^iX \infty {}^i d & \text{is} & (\lambda)^d V - \\ {}^i M \infty {}^i X & \text{is} & (\lambda)^d V \\ {}^i M \infty {}^i d & \text{is} & 0 \end{array} \right\} = {}^v \nabla$$

Notons que $\lambda_j(x)$ est une quantité non négative et que les conditions de l'exposition (3.9) peuvent être réécrites de la façon suivante, sous la forme $d_i = M_i W_i$, $\lambda_j = \bar{\lambda}$ et $d_i = \lambda_j W_i$ pour tout $i = 1, \dots, N$. Ceci

Exemple 3.1 Pour un plan d'échantillonnage en grappes à

Nous pouvons obtenir le même résultat que celui de l'exemple 3.1 au moyen de $Y = M \cdot Y$. Il s'agit de l'estimateur par le quotient, qu'on peut utiliser si l'on connaît M . La situation où les poids d'échantillonnage globaux sont constants pour tous les éléments (c'est-à-dire plan d'échantillonnage autopondéré) est un cas particulier bien connu. Nous y reviendrons à la section 4.

Exemple 3.2 Plan d'échantillonnage aléatoire simple en grappes à un degré no plan d'échantillonnage à deux degrés nous avons $w_{j|I} = c_j$, et $p_i^* = 1/N$ pour tous $i \in J$, et il découle donc de (3.8) et de (3.9) que $\Delta^0 = 0$ et

approximativement à $m \cdot CV_M^2 / CV_y^2$ quand $N/D_i = 1$

L'exemple 3.2 montre que, si l'inégalité des tailles de grappe n'est pas reflétée dans le plan d'échantillonnage, l'efficacité relative de \bar{Y} par rapport à \bar{y} dépend partiellement de la variabilité relative de ces tailles de grappe. Si les moyennes de grappe sont toutes égales, alors l'échantillonnage en grappes rend \bar{Y} plus efficace que \bar{y} , et inversement, si tous les totaux de grappe sont égaux. Par ailleurs, si toutes les grappes sont de même taille, l'échantillonnage

la variance d'échantillon relative des poids. Strictement parlant, cette formule n'est pas indépendante de la variable y , parce que p dépend de cette variable. En outre, l'effet de plan pourrait ne pas être exempt de l'unité de mesure, à moins que $V_p(\hat{Y})$ soit exprimé sous une forme factorielle de S_y^2/m . Voir Park et Lee (2002). La formule (2.2) n'est valide que s'il n'existe aucune corrélation entre les poids d'échantillonage et la variable étudiée y . Par contre, s'il existe une corrélation, il peut être nécessaire de modifier la formule conformément aux études de Spencer (2000) et de Park et Lee (2001). À la section suivante, nous exposons cet aspect en détail pour l'échantillonage à deux degrés et nous examinons aussi ce point plus en profondeur à la section 4.1.

3. DÉCOMPOSITION DE L'EFFET DE PLAN

Sous ÉCHANTILLONAGE À DEUX DEGRÉS

Considérons un plan d'échantillonage réalisé en deux degrés. Supposons qu'une population $U = \{k: k = 1, \dots, M\}$ avec M éléments soit regroupée en N grappes de taille M_j , telle que $M = \sum_{j=1}^N M_j$. L'échantillon de premier degré $s_a = \{i: i = 1, \dots, n\}$ de n grappes (unités primaires d'échantillonage ou UPE en abrégé) est tiré avec remise à partir de N grappes avec probabilité p_i , où $\sum_{i=1}^N p_i = 1$. Soit $p_a = \text{Pr}(s_a)$ le plan d'échantillonage de premier degré. L'échantillon de deuxième degré $s_{ji} = \{j: j = 1, \dots, m_j\}$ de m_j éléments (unités secondaires d'échantillonage ou USE) m_j éléments tiré indépendamment à partir de chaque UPE i sélectionnée à la première étape selon un plan d'échantillonage arbitraire, disons $p_{ji} = \text{Pr}(s_{ji}|s_a)$, où $i \in s_a$. Représentons l'échantillon total d'éléments et le plan d'échantillonage global par $s = \cup_{i \in s_a} s_{ji}$ et $p = \text{Pr}(s)$, respectivement. Au j^{e} élément de la i^{e} grappe est associée une caractéristique d'enquête y_{ji} , $j = 1, \dots, M_j$, $i = 1, \dots, N$. Pour un $i \in s_a$, donnés, représentons par w_{ji} les poids d'échantillonage de deuxième degré tels qu'un estimateur de la forme $\hat{Y}_i = \sum_{j=1}^{m_j} w_{ji} y_{ji}$ soit sans biais pour le total de grappe $X_i = \sum_{j=1}^{m_j} y_{ji}$, c'est-à-dire $E_p(\hat{Y}_i) = X_i$, où E_p représente l'espérance par rapport à l'échantillonage de deuxième degré. Soit $w_i = 1/(n p_i)$ les poids d'échantillonage de deuxième degré. Soit $X = \sum_{i=1}^N X_i$ le total de population. Il est facile de montrer que $E_p(X/p) = X$. Supposons que X_i est connu pour $i \in s_a$, $\sum_{i=1}^N w_i X_i$ est la moyenne des n estimateurs sans biais de X , où $E_a(\sum_{i=1}^n w_i X_i) = X$, où E_a représente l'espérance par rapport au plan d'échantillonage (une grappe ou un élément) soit sélectionné plus d'une fois, mais soit traité différemment. Définissons les expressions

(3.4)
$$Y_{\text{cas}} = M \bar{Y}_{\text{cas}} = \frac{f}{I} \sum_{i=1}^I Y_i$$

serviraient d'estimateurs de la moyenne \bar{Y} et du total X de la population, respectivement, sous casar, où $f = m/M$ est la fraction d'échantillonage globale. Leur variance sous ce plan d'échantillonage est donnée par $V_{\text{casar}}(Y_{\text{cas}}) = M^2 V_{\text{casar}}(\bar{Y}_{\text{cas}})$, où $V_{\text{casar}}(\bar{Y}_{\text{cas}}) = m^{-1} S_y^2$ et $S_y^2 = (M-1)^{-1} \sum_{i=1}^M (Y_i - \bar{Y})^2$. Nous notons que m est la taille d'échantillon obtenue, qui, en général, est une quantité aléatoire. À partir de (3.1), de (3.3) et des expressions susmentionnées, en remplaçant m par sa valeur prévue m' par rapport au plan d'échantillonage global p , c'est-à-dire $m' = E_p(\hat{m})$, nous pouvons représenter les effets de plan pour Y et X par les expressions

(3.3)
$$A V_p(\hat{Y}) = \frac{1}{M^2} \left[\sum_{i=1}^N w_i^2 X_i^2 + \sum_{i=1}^N w_i X_i Y_i \right]$$

Si nous tirons un échantillon aléatoire simple de taille $m = \sum_{i=1}^N m_i$ avec remise à partir de la population U , alors une moyenne d'échantillon $\bar{y}_{\text{sts}} = \sum_{i=1}^m y_i/m$ et son développement

(3.2)
$$\bar{Y} \approx \bar{y} + M^{-1} D$$

où $\bar{D} = \sum_{i=1}^N \sum_{j=1}^{m_j} w_{ji} d_{ji}$ est un estimateur sans biais du total de population $D = \sum_{i=1}^N \sum_{j=1}^{m_j} d_{ji}$ de $d_{ji} = y_{ji} - \bar{Y}$, qui représente l'écart de y_{ji} par rapport à la moyenne de population \bar{Y} . Notons que $D = 0$. En écrivant $D_i = \sum_{j=1}^{m_j} d_{ji}$, $Y_i - M_i \bar{Y}$ et $D_i = \sum_{j=1}^{m_j} w_{ji} d_{ji}$, nous obtenons la variance approximative de \bar{Y} pour l'expression (3.2) sous la forme

(3.1)
$$V_p(\hat{Y}) = V_a(E_p(\hat{Y})) + E_a(V_p(\hat{Y}))$$

Y peut s'écrire sous la forme

Un estimateur de la moyenne de population $\bar{Y} = X/M$ utilisé fréquemment est l'estimateur (par le quotient) pondéré donné par $\hat{Y} = X/M$, où $M = \sum_{i=1}^N \sum_{j=1}^{m_j} w_{ji}$. En recourant à la linéarisation de Taylor, telle que décrite dans Särndal et coll. (1992, pages 176–178), nous pouvons approximer \hat{Y} par

pages 151–152).

dégré et de deuxième degré. Voir Särndal et coll. (1992, pages 151–152).

où V_p , V_a et V_b représentent les variances définies par rapport aux plans d'échantillonage global, de premier degré et de deuxième degré.

$$Deff = \frac{V_p(\hat{\bar{y}})}{(1-f)S_y^2/m} \quad (2.1)$$

où V_p est la variance par rapport à p , $f = m/M$ est la fraction d'échantillonnage globale et $S_y^2 = (M-1)^{-1} \sum_{i=1}^M (y_i - \bar{Y})^2$ est la variance d'élément de la population de la variable y . Bien qu'au départ, l'effet de plan ait été défini pour un estimateur de la moyenne de population (Kish 1995), on peut le définir pour toute statistique significative calculée d'après un échantillon tiré selon un plan d'échantillonnage complexe.

Le Deff est une quantité de population qui dépend de plan d'échantillonnage et se rapporte à une statistique particulière estimant un paramètre de population d'intérêt donné. Divers estimateurs permettant d'estimer un même paramètre ont des effets de plan différents, même sous des plans d'échantillonnage identiques. Par conséquent, l'effet de plan englobe non seulement l'efficacité de plan d'échantillonnage, mais aussi celle de l'estimateur. Särndal, Swensson et Wretling (1992, page 54) ont bien établi ce point en définissant l'effet de plan comme une fonction de plan de sondage (p) et de l'estimateur ($\hat{\theta}$) du paramètre de population ($\theta = \theta(y)$). Donc, nous pouvons l'écrire sous la forme

$$Deff_p(\hat{\theta}) = \frac{V_p(\hat{\theta})}{V_p^{censar}(\hat{\theta})}$$

où $\hat{\theta}$ est la forme habituelle d'un estimateur de θ sous eassr, qui est normalement différent de θ . Par exemple, pour estimer la moyenne de population, nous pourrions utiliser la moyenne (quotient) pondérée $\hat{\theta} = \sum w_k y_k / \sum w_k$ avec les poids d'échantillonnage w_k , tandis que θ serait la moyenne d'échantillon simple $\sum y_k / m$, où la sommation est faite sur l'échantillon s . Nous examinerons l'incidence d'estimateurs particuliers sur l'effet de plan plus loin.

Plus tard, Kish (1995) a préconisé l'utilisation d'un paramètre défini de façon légèrement différente, qu'il a appelé Deff et dont le dénominateur est la variance sous échantillonnage aléatoire simple avec remise (eassar), étant donné que l'échantillonnage sans remise fait partie de plan de sondage et devrait donc être reflété dans la définition. Kish a également argumenté que le paramètre Deff est plus facile à utiliser pour faire des inférences et qu'il vaut mieux définir l'effet de plan sans le facteur de correction pour population finie $(1-f)$, car ce dernier est difficile à calculer dans certaines situations. La nouvelle définition est donnée par

$$Deff_p(\hat{\theta}) = \sqrt{\frac{V_p(\hat{\theta})}{V_p^{censar}(\hat{\theta})}}$$

ou $Deff_p^2(\hat{\theta}) = V_p(\hat{\theta}) / V_p^{censar}(\hat{\theta})$. Les logiciels applicables à des données d'enquête, tels que WesVar et SUDAAN, produisent $Deff^2$ au lieu de Deff. Nous utiliserons cette

définition dans le présent article. Quand le paramètre de population est le total (Y), l'estimateur sans biais est le total pondéré d'échantillon, c'est-à-dire $\hat{Y} = \sum w_k y_k$. Si le paramètre d'intérêt est la moyenne de population, on l'estime habituellement par la moyenne pondérée, qui est $\hat{\bar{Y}} = \sum w_k y_k / \sum w_k$. Il s'agit d'un cas particulier de l'estimateur par le quotient, $\sum w_k y_k / \sum w_k x_k$, où $x_k \equiv 1$ pour tout $k \in s$.

Une idée fausse courante concernant les effets de plan pour \bar{Y} et $\hat{\bar{Y}}$ est que leurs valeurs sont les mêmes. Or, on a constaté que l'effet de plan pour \bar{Y} , $Deff_p^2(\bar{Y})$, a tendance à être beaucoup plus grand que celui pour $\hat{\bar{Y}}$, $Deff_p^2(\hat{\bar{Y}})$. Cette différence a aussi été mentionnée, entre autres, par Kish (1987) et par Barron et Finch (1978). Une explication est donnée par Hansen et coll. (1953, volume I, pages 336–340) qui montrent que la différence a pour origine la variance relative des tailles de grappe. Plus récemment, Särndal et coll. (1992, pages 315–318) ont montré que, contrairement au cas de \bar{Y} , l'effet de plan pour $\hat{\bar{Y}}$ dépend de la variation (relative) de la variable y . En fait, même l'effet de plan pour $\hat{\bar{Y}}$ peut dépendre de la variation (relative) de la variable y , ce dont nous discuterons à la section 4. Cette dépendance est en contradiction avec ce que l'effet de plan est destiné à mesurer comme Kish (1995) l'a décrit explicitement :

[Traduction] « Deff est utilisé pour exprimer les effets de plan d'échantillonnage au-delà de la variabilité élémentaire (S_y^2/m), en éliminant à la fois les paramètres de nuisance que sont les unités de mesure et la taille d'échantillon. Si l'on élimine S_y , les unités et la taille d'échantillon m , les effets de plan sur les erreurs d'échantillonnage deviennent généralisables (transférables) à d'autres statistiques et à d'autres variables, dans la même enquête, voire même dans d'autres. »

Sa déclaration peut être vaguement correcte pour la moyenne pondérée $\hat{\bar{Y}}$ telle qu'exprimée dans la formule approximative d'échantillon fréquemment utilisée de $Deff_p^2(p, \hat{\bar{Y}})$ donnée par Kish (1987) :

$$Deff_p^2(\hat{\bar{Y}}) = \{1 + p(m-1)\} \{1 + cv_w^2\} \quad (2.2)$$

où le plan d'échantillonnage p contient des caractéristiques complexes, telles qu'une pondération inégale et un échantillonnage en grappes, $p = p(y)$ est le coefficient de corrélation intraclasse (souvent appelé mesure de l'homogénéité à l'intérieur des grappes), m est la taille

Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe

INHO PARK et HYUNSHIK LEE¹

RÉSUMÉ

Nous examinons de nouveau la relation entre les effets de plan pour l'estimateur pondéré du total et l'estimateur pondéré de la moyenne sous échantillonnage complexe. Nous donnons des exemples sous diverses conditions. En outre, au moyen d'exemples, nous corrigeons certaines idées fausses concernant les effets de plan.

MOTS CLÉS : Échantillonnage aléatoire simple; échantillonnage ppi; échantillonnage à plusieurs degrés; autopondération; stratification a posteriori; coefficient de corrélation intragroupe.

1. INTRODUCTION

L'effet de plan est un paramètre dont l'usage est très répandu en échantillonnage pour élaborer le plan d'échan-

lillonnage et indiquer l'effet de ce dernier sur l'estimation et l'analyse. On le définit comme étant le rapport entre la variance d'un estimateur sous échantillonnage complexe et celle d'un estimateur sous échantillonnage aléatoire simple pour la même taille d'échantillon. Les projections applicables aux enquêtes complexes, tels que WesVar et SUDAAN, produisent systématiquement une estimation de l'effet de plan. Au départ, on a défini ce dernier en vue de calculer l'estimateur (par le quotient) pondéré de la moyenne de la population (Kish 1995). Cependant, en pratique, il est courant d'appliquer ce concept à d'autres statistiques, comme l'estimateur pondéré du total de population, souvent avec fruit, mais parfois avec confusion et en comprenant mal le problème. Cette dernière situation se présente surtout lorsqu'on applique des résultats simples, mais utiles, obtenus au moyen d'un plan d'échantillonnage assez simple, à des problèmes plus complexes. Dans le présent article, nous examinons la relation entre les effets de plan pour l'estimateur pondéré du total et l'estimateur pondéré de la moyenne sous divers plans d'échantillonnage complexes. À la section 2, nous passons brièvement en revue la définition de l'effet de plan et son utilisation pratique, et nous discutons de certaines idées fausses sur les effets de plan de sondage dans le calcul des estimateurs pondérés du total et de la moyenne. À la section 3, nous analysons la différence entre les effets de plan pour l'estimateur pondéré du total et pour celui de la moyenne sous un plan d'échantillonnage à deux degrés. Puis, à la section 4, nous discutons des effets de plan sous divers plans d'échantillonnage à deux degrés et certains cas plus généraux, et nous essayons de corriger certaines idées fausses au moyen

2. BRÈVE REVUE DE LA DÉFINITION ET DE L'UTILISATION DE L'EFFET DE PLAN EN PRATIQUE

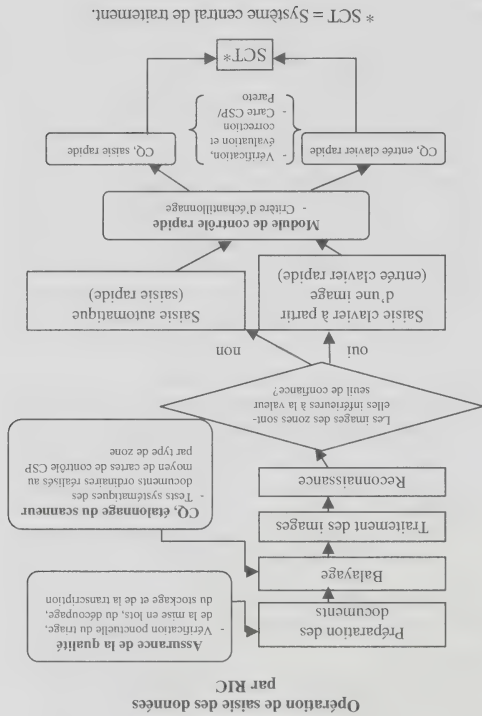
de ces exemples. Enfin, à la section 5, nous résumons notre discussion.

Un précurseur de l'effet de plan popularisé par Kish (1965) a été utilisé par Cornfield (1951). Ce dernier a défini l'efficacité d'un plan d'échantillonnage complexe pour l'estimation d'une proportion de population comme étant le rapport de la variance de l'estimateur de la proportion sous échantillonnage aléatoire simple avec remise (casar) à la variance correspondante sous échantillonnage aléatoire simple en grappes pour la même taille d'échantillon. L'inverse du rapport défini par Cornfield (1951) a également été utilisé par d'autres auteurs. Ainsi, Hansen, Hurwitz et Madow (1953, volume I, pages 259-270) discutent de l'augmentation de la variance relative d'un estimateur par le quotient due à l'effet de mise en grappes de l'échantillonnage en grappes par rapport à l'échantillonnage aléatoire simple sans remise (casar). Toutefois, l'expression effet de plan, ou Deff en abrégé, a été inventée et définie officiellement par Kish (1965, section 8.2, page 258) comme étant « le rapport de la variance réelle d'un échantillon à la variance d'un échantillon aléatoire simple contenant le même nombre d'éléments » [traduction] (pour un historique pour détaillé, voir aussi Kish 1995, page 73 et les références mentionnées dans ce texte). Supposons que nous voulions estimer la moyenne de la population (\bar{Y}) d'une variable y d'après un échantillon de taille m tiré selon un plan d'échantillonnage complexe représenté par p dans une population de taille M . Le Deff de Kish d'une estimation (\bar{y}_p) est donné par

REMERCIEMENTS

Les auteurs remercient le rédacteur en chef, un rédacteur associé et un rédacteur adjoint de leurs commentaires détaillés et constructifs. Ils remercient aussi Bob Bougie pour ses nombreux commentaires utiles.

ANNEXE



5. CONCLUSION

Il est clair, d'après les résultats de la présente analyse, que les procédures d'AQ/CQ ont été fort utiles et ont eu un effet très constructif sur l'ensemble de l'opération. Les procédures d'AQ appliquées au processus de préparation des documents ont permis d'empêcher un grand nombre de documents de mauvaise qualité d'atteindre les postes de balayage et d'étiqueter ceux qui l'ont fait afin de leur faire subir un traitement spécial et une vérification totale subséquente.

Les procédures de CQ ont ensuite permis d'optimiser le réglage des machines grâce au processus de vérification de l'étalonnage avant la production. En outre, durant la production, le tirage d'échantillons de CQ a également permis de repérer certains problèmes posés par les processus de reconnaissance automatique et de saisie clavier à partir d'une image, de sorte qu'ils ont pu être corrigés au besoin. Dans tous les cas, des mesures objectives ont fourni des signaux d'avertissement précoces, à chaque étape du traitement, et des mesures correctives et préventives ont été prises, au besoin. Une rétroaction systématiquement importante des étapes du processus de RIC a permis d'améliorer toutes les étapes du processus de RIC à perpétuer la qualité.

BIBLIOGRAPHIE

- KALPIC, D. (1994). Miscellaneous, Automated Coding of Census Data *Journal of Official Statistics*, 10, 4, 449-463.
- MUDRYK, W., BOUGIE, B. et XIE, H. (2001). Quality Control of ICR Data Capture: 2001 Canadian Census of Agriculture, *International Conference on Quality in Official Statistics in Stockholm, Sweden*.
- MUDRYK, W., et XIE, H. (2002). Quality Control Application in ICR Data Capture for the 2001 Canadian Census of Agriculture, *Proceedings of the Joint Statistical Meetings, American Statistical Association*, 2424-2429.
- PASLEY, B. (2000). Exclusivement sur le Web: The Good and Bad of Scanned Images. Affiché à l'entrée sur le site.

Les résultats du CQ montrent clairement que les documents *aberrants* ont un effet négatif plus important sur le processus de *saisie rapide* (taux d'erreurs de substitution de 7,0 %) que sur le processus d'entrée clavier rapide (taux d'erreurs de saisie de 3,7 %). Par conséquent, le processus de filtration en vue de réserver un traitement particulier aux documents *aberrants* était une importante mesure à prendre. Les résultats du CQ montrent aussi que, si les questionnaires

processus de saisie *rapide* et 23 %, grâce au processus qu'environ 77 % des zones ont été saisies grâce au leur taux d'erreurs plus élevé. Dans l'ensemble, nous traitons les cas les plus difficiles, ce qui explique en partie connaître les opérateurs, il faut toutefois reconnaître qu'ils donnent des saisies (c'est-à-dire 2,0 % c. 3,4 %). En ce qui concerne les saisies, il faut souligner que les données sont moins nombreuses que les données de saisie *à partir d'une image*. Cette observation est assez importante, puisqu'elle sous-entend des économies manifestes associées d'une amélioration de la qualité des données saisies (c'est-à-dire 2,0 % c. 3,4 %). En ce qui

termiination des causes profondes de nombreux problèmes et à leur résolution. Ces informations rétroactives ont permis d'améliorer continuellement la qualité aux diverses étapes du processus.

Pour une description détaillée de ces résultats d'AQ/CQ, consulter Mundryk et Xie (2002).

Nombre estimé de zones vérifiées et d'erreurs avant le contrôle	AOQ (%)
1 70 249	1,99
1 044 358	0,00
1 214 607	1,90

2,16	3,24	2 135 594
0,00	5,09	1 731 092
2,29	2,82	404 502

Tableau I

Saisie rapide		Processus		de questionnaires dans la population		de zones dans la population		de zones vérifiées et d'erreurs avant le contrôle		AOQ (%)	
Questionnaires ordinaires	273 818	21 248 277	1 702 249	2 01	1,99	Questionnaires ordinaires	286 520	22 292 635	1 214 607	2,95	1,90
Questionnaires aberrants	12 702	1 044 358	1 044 358	6,99	0,00	Questionnaires aberrants	25 788	686 734	3,67	0,00	2,97
Globalement						Globalement					
Entrée clavier rapide						Questionnaires ordinaires	281 502	6 376 020	3,41	3,28	
Questionnaires ordinaires	27 624 297	1 731 092	29 355 389	2 135 594	3,24	Questionnaires ordinaires	27 624 297	1 731 092	2,82	2,29	
Questionnaires aberrants	1 731 092	29 355 389	2 135 594	3,24	2,16	Questionnaires aberrants	25 788	686 734	3,67	0,00	
Globalement						Globalement					

4.5 Sommaire du CQ

4.5 Sommaire du CQ

processus de *saisie rapide* et d'entrée clavier *rapide*. Nous constatons que l'AQO globale est de 1,90 % pour le processus de *saisie rapide* et de 2,97 % pour le processus d'entrée clavier *rapide*. Ces chiffres représentent une baisse importante comparativement aux estimations correspondantes de la qualité avant le contrôle de 2,95 % et de 3,45 %, respectivement. L'estimation de l'AQO globale pour les deux processus est de 2,16 % (comparativement à une qualité globale avant le contrôle de 3,24 %). Il convient de souligner que nous avons supposé que l'AQO pour les questionnaires *aberrants* était égale à 0 %, puisque tous les documents *aberrants* ont été subséquemment vérifiés entièrement.

4.2 Vérification de l'étalonnage du scanner

Afin d'assurer le réglage et l'étalonnage optimal des scanners, nous avons procédé à une *vérification de l'étalonnage*, au départ, de dix par jour et, par après, une fois par jour, avant le traitement de la production. De nombreux lots d'essai balayés durant l'opération ont donné un taux de rejet assez élevé pour chaque scanner. En moyenne, environ deux à trois essais par jour (avec les réévaluations correspondantes) ont été nécessaires pour optimiser le réglage de chaque machine. Ces résultats montrent qu'il est nécessaire de réétalonner les scanners entre les cycles de traitement. Il convient de souligner que certains rejets étaient dus à des problèmes particuliers au lot d'essai qui ont été corrigés plus tard. Il s'agit définitivement d'un domaine où certaines améliorations de procédure seront nécessaires dans l'avenir.

Durant ce test, les deux scanners ont manifesté une variabilité raisonnablement élevée. Le nombre élevé d'essais requis, le taux élevé de rejets et la forte variabilité au cours des divers cycles de traitement pour nombre de types de zones prouvent qu'il est nécessaire d'étalonner l'appareil de balayage correctement avant la production. Simon, les scanners pourraient être accidentellement réglés de façon à produire des images médiocres dès le début, ce qui rendrait fort difficile une saisie de bonne qualité. Quant un lot d'essai était rejeté, en général, nous avons déterminé quel était le problème et puis ensuite des mesures d'entretien et de correction. Ces mesures incluaient, par exemple, la reconfiguration du scanner, le remplacement des ampoules électriques usées, la résolution des problèmes logiciels, le nettoyage des têtes de lecture sales, etc. Grâce à ce test, nous avons pu étalonner les scanners et maintenir un niveau élevé de rendement entre les cycles de production.

4.3 Saisie rapide et entrée clavier rapide

Pour le processus de *saisie rapide*, au cours des 18 semaines complètes de traitement des questionnaires

Afin d'assurer le réglage et l'étalonnage opti-

Durant ce test, les deux scanners ont manifesté une variabilité raisonnablement élevée. Le nombre élevé d'essais requis, le taux élevé de rejets et la forte variabilité au cours des divers cycles de traitement pour un nombre de zones prouvent qu'il est nécessaire d'étalonner l'appareil de balayage correctement avant la production. Sinon, les scanners pourraient être accidentellement réglés de façon à produire des images médiocres dès le début, ce qui rendrait fort difficile une saisie de bonne qualité. Quant un lot d'essai était rejeté, en général, nous avons déterminé quel était le problème et pris ensuite des mesures d'entretien et de correction. Ces mesures incluaient, par exemple, la reconfiguration du scanner, le remplacement des ampoules électriques usées, la résolution des problèmes logiciels, le nettoyage des têtes de lecture, *etc.* Grâce à ce test, nous avons pu étalonner les scanners et maintenir un niveau élevé de rendement entre les cycles de production.

Pour le processus de *saisie rapide*, au c

4.3 Saisie rapide et entrée clavier rapide

Pour le processus de *saisie rapide*, au cours des 18 semaines complètes de traitement des questionnaires

L'objectif p

4.4 Estimations de la qualité moyenne après le contrôle

05/05/97 22:05:00 05/05/97 22:05:00 05/05/97 22:05:00

4.4 Estimations de la qualité moyenne après le contrôle

d'accepter ou de rejeter l'échantillon. Le taux de rejet serait calculé pour une carte de contrôle p standard. Quand le taux d'erreurs dans l'échantillon était supérieur à ce niveau, le processus a été rejeté et l'examineur du CQ a recherché et appliqué des mesures correctives, au besoin; sinon, le processus a été accepté.

Nous avons procédé à l'échantillonnage en nous basant sur un scanneur individuel pour la saisie rapide et sur un opérateur individuel pour l'entrée clavier rapide. De temps à autre, le nombre de questionnaires à échantillonner était plus élevé pour certains opérateurs que pour d'autres, selon leur rendement réel. Puisque les observations réelles étaient fondées sur des échantillons, une variabilité ordinaire de $+3\sigma$ a été permise en sus de la norme de qualité prévue (c'est-à-dire la ligne centrale d'une carte de contrôle p). Pour chaque processus, pour ces observations échantillonnées, les décisions concernant les lots ont été faites par le système durant la vérification CQ, puis les résultats ont été portés par après sur une carte de contrôle p pour chaque scanneur et chaque opérateur, et les cartes de contrôle ont été mises à jour hebdomadairement.

Pour une description détaillée de ces procédures d'AQ/CQ et de leur justification, consulter la Mudryk, Bougie et Xie (2001).

3. AMÉLIORATIONS DE LA QUALITÉ

Pour l'opération de saisie des données par RLC, nous avons inclus deux éléments essentiels dans la stratégie d'amélioration de la qualité. Ces éléments sont une réaction des résultats de l'AQ/CQ et l'application de mesures correctives et préventives, au besoin. Ces deux éléments ont permis à divers membres du personnel de jouer un rôle actif dans l'amélioration de la qualité de chaque processus grâce à une meilleure compréhension des problèmes déistes et aux mesures correctives et préventives subséquentes qui ont été prises.

En nous appuyant sur l'analyse des données CQ, nous avons examiné tous les processus afin de déterminer s'ils se déroulaient efficacement. Nous avons tenu des réunions de CQ hebdomadaires avec les employés des opérations afin de passer en revue les progrès concernant l'opération complète. Durant ces réunions, nous avons abordé les problèmes ayant une incidence sur tout processus et fait des recommandations afin d'en traiter les causes profondes et d'empêcher qu'ils se reproduisent. La participation des employés des opérations à la résolution de ces problèmes a grandement facilité l'amélioration continue de la qualité. Les exemples plus importants prises durant l'opération qui ont abouti à des améliorations de la qualité à diverses étapes.

Exemple 1 : Processus de filtrage pour la détection des documents aberrants

Durant les premières semaines de production, nous avons constaté que certains documents provoquaient un grand nombre d'erreurs à cause d'une grande croix en travers d'une page, de 0 ou de tirets dans diverses zones, etc. Ces documents donnaient un taux d'erreurs élevé lors des formes de saisie, mais surtout dans le processus de saisie mécanique rapide. Puisque ces documents étaient très différents de la majorité des documents ordinaires, nous avons introduit une procédure pour les tirer en vue d'un traitement spécial ultérieur. En fait, certains documents ont dû être retranscrits à cette étape avant de pouvoir être traités par RLC.

Exemple 2 : Ajustement des réglages du système pour le balayage et la reconnaissance

Les faits saillants des sommaires hebdomadaires de CQ indiquaient que les deux scanneurs faisaient fréquemment des erreurs aux pages 3 et 14 des questionnaires durant les premières semaines de traitement. Une enquête nous a permis de constater qu'il existait un problème de lecture du modèle de la page 3 et que le seuil préétabli de reconnaissance pour les zones numériques de la page 14 était trop faible. Après correction des réglages du système sur les deux scanneurs, nous avons observé une amélioration considérable du balayage de ces deux pages.

Exemple 3 : Nouvelle formation des opérateurs dont le taux d'erreurs était élevé

Durant l'opération d'entrée clavier, les résultats de CQ ont indiqué que le processus de « saisie à partir d'une image » causait plus de difficultés que la normale à certains opérateurs et que le taux d'erreurs de ces derniers restait élevé pendant plusieurs semaines. Dans le contexte d'une amélioration continue de la qualité, nous avons offert une nouvelle formation régulière à ces opérateurs. Cette mesure a permis à nombre d'entre eux d'améliorer significativement (semaine après semaine) leur rendement d'entrée clavier.

4. ÉVALUATION ET ANALYSE DU CQ

Tout au long de l'opération, nous avons produit de nombreux rapports, graphiques et estimations de CQ afin de fournir des renseignements sur les niveaux de qualité avant et après le contrôle et d'évaluer la sortie de chaque processus de production. Nous avons utilisé ces rapports pour analyser la qualité hebdomadaire de chaque processus, ainsi que les variations d'une semaine à l'autre.

4.1 Préparation des documents

Nous avons appliqué aux cinq sous-processus de la préparation des documents des procédures individuelles d'AQ à diverses fréquences et pris continuellement des mesures correctives ainsi que préventives, comme le dicte la loi. Nous avons appliqué aux cinq sous-processus de la loi d'AQ ont permis d'améliorer

machine et l'obtention de bons résultats lors d'un nouveau test. À l'étape du balayage, la reconnaissance mécanique des caractères peut aboutir à la substitution de valeurs incorrectes si les images produites sont de mauvaise qualité. De telles images pourraient être dues à de nombreux facteurs, tels que l'encreusement des têtes de lecture, la saisie des fenêtres optiques, un mauvais alignement, un mauvais enregistrement des zones, un mauvais niveau de contraste/brillance, des problèmes d'alimentation du papier, etc. Puisqu'une norme particulière de qualité a été établie pour chaque type de zone, nous avons utilisé une carte de contrôle p distincte pour évaluer le taux d'erreurs de substitution pour chaque type (plus précisément, alphabétique, alphanumérique, numérique, cases à cocher et code à barres). Puisque la norme de qualité acceptable avait été établie a priori pour chaque type de zone par le programme client, nous avons utilisé comme mesure de qualité le « pourcentage de zones erronées », c'est-à-dire le taux d'erreurs de substitution selon le type de zone pour chaque scaneur.

En nous fondant sur la théorie des cartes de contrôle du CSP, nous avons suivi les règles qui suivent pour prendre une décision pour chaque test d'étalonnage du scaneur :

- Si les taux d'erreurs dans l'échantillon pour les cinq types de zone étaient tous inférieurs à la limite supérieure de contrôle (LSC), nous avons conclu que le système de balayage fonctionnait correctement et était prêt pour le balayage de production.
- Sinon, nous avons conclu que le scaneur posait un problème et que des mesures correctives devaient être prises avant de lancer la production régulière.

Les lots d'essai ont été créés en cherchant à réduire au minimum la taille d'échantillon pour chaque type de zone, de telle sorte que le niveau de confiance du producteur soit au moins de 95 %. Puis, nous avons utilisé ce chiffre comme guide pour sélectionner les questionnaires réels pour chaque lot d'essai. La taille minimale était requise pour chaque type de zone afin d'atteindre le haut niveau d'efficacité des décisions requis pour l'essai d'étalonnage du scaneur, tandis que le niveau de confiance du producteur renvoyait à la vraisemblance que le balayage ne soit pas rejeté lors du test pour le type de zone en question lorsque le système fonctionnait au niveau cible acceptable. Pour chaque type de zone, nous avons calculé la limite supérieure de contrôle en supposant une variabilité de $+2\sigma$. Cette limite est inférieure à la limite supérieure de contrôle habituelle de $+3\sigma$, puisque la vérification de l'étalonnage du scaneur a été conçue de façon à être plus sensible aux décalages plus petits au démarrage que durant la production normale.

2.2.3 Saisie rapide et entrée clavier rapide

Après le balayage des questionnaires, le système produit une image numérique de chaque zone ainsi qu'une interprétation de sa valeur et un niveau de confiance

zone confondus « pourcentage de zones erronées » pour tous les types de processus, la mesure de qualité a été définie comme étant le pourcentage de zones erronées. Pour les deux claviers pour évaluer ce processus manuel. Pour les deux claviers pour évaluer ce processus automatique, nous avons utilisé les taux d'erreurs de formation, la fatigue, etc., par exemple, une mauvaise formation, la fatigue, etc., par exemple, pour de nombreuses raisons, dont le manque de compétence, les opérateurs peuvent faire des erreurs de frappe rapide, les opérateurs peuvent faire des erreurs de frappe longue période. Dans le cas de l'opération d'entrée clavier de reconnaissance pour de nombreuses zones pendant une longue période, les opérateurs peuvent faire des erreurs de frappe si elles ne sont pas corrigées, elles peuvent influencer le taux de reconnaissance. Ces erreurs de substitution sont particulièrement graves, car, nous avons utilisé les taux de substitution pour évaluer ce processus, pour la zone en question et, par conséquent, nous avons machine produise une valeur différente de la valeur réelle dans le cas de l'opération de saisie rapide, il arrive que la qualité prévue pour le processus en question.

Dans le cadre des deux opérations de saisie, nous avons distingué deux catégories de traitement des documents basées sur les questionnaires ordinaires et les questionnaires aberrants. Des procédures de CQ ont été mises en place pour chaque catégorie. Un échantillon distinct a été nécessaire pour chaque processus, un pour la saisie rapide et un pour l'entrée clavier rapide. Le système pouvait faire la distinction entre les zones de saisie rapide et celles de l'entrée clavier rapide dans chaque questionnaire échantillon et tenir des comptes distincts des zones saisies selon chaque processus. Ces nombres de zones sont devenus, en dernière analyse, la taille d'échantillon pour chaque échantillon. Nous avons ensuite comparé chaque échantillon à son propre taux de rejet seuil, qui était une fonction du nombre de zones observées (c'est-à-dire la taille d'échantillon effective) et de la norme de qualité ou cible prévue pour le processus en question. La décision a alors été prise

l'opération. Les objectifs particuliers de chaque étape étaient les suivants :

- a) Préparation des documents : veiller à ce que seuls des documents lisibles passent à l'étape du balayage.
- b) Etalonnage du scanner : veiller à ce que le réglage et l'étalonnage de la machine soient optimaux avant le début de la production.
- c) Saisie rapide (saisie mécanique) et entrée clavier rapide (saisie manuelle) : veiller à ce que la qualité de la saisie des données durant la production soit de haut niveau.

2.2 Méthodes d'AQ/CQ

Les diverses étapes importantes du traitement étaient opérationnellement uniques en leur genre, les exigences en matière de qualité différaient. Par conséquent, nous avons appliqué des procédures d'AQ à l'opération de préparation des documents et des procédures de CQ aux opérations d'étalonnage du scanner, de saisie rapide et d'entrée clavier rapide. Un ordigramme qui montre les diverses étapes de l'opération de saisie des données par RIC et les endroits exacts où les procédures d'AQ et de CQ ont été appliquées figure en annexe.

2.2.1 Préparation des documents

L'opération de préparation des documents a, essentiellement, été subdivisée en cinq sous-processus, à savoir le triage, la transcription, la mise en lots, le découpage et le stockage. Cette opération, qui avait pour but de préparer les questionnaires et les lots connexes pour le balayage par l'appareil de RIC, a été exécutée manuellement par le personnel de bureau. Elle comportait des activités telles que le triage du contenu des enveloppes reçues selon le type de document (*triage*), la retranscription des questionnaires endommagés ou illisibles (*transcription*), le groupement des questionnaires en lots pour l'entregistrement (*mise en lots*), le découpage du dos de chaque cahier questionnaire au moyen d'un couteau électronique (*déconvolage*) et le classement des questionnaires dans les archives (*stockage*). L'un des aspects les plus importants de cette opération était le réglage et l'isolement des questionnaires barrés d'une croix ou de questionnaires présentant des écritures sur les zones de données, des marques inappropriées ou des entrées illisibles, ou de questionnaires déchirés, chiffonnés ou collés au moyen de

ruban adhésif, etc.

Les erreurs susceptibles de survenir durant cette opération peuvent causer des problèmes à l'étape du balayage. Nous avons estimé que des procédures d'AQ seraient appropriées pour assurer la qualité à cette étape, puisque le nombre de fonctions administratives étaient, elles aussi, soumises à diverses vérifications croisées automatisées du

2.2.2 Vérification de l'étalonnage du scanner

L'expérience a montré que, si le scanner n'est pas configuré correctement, le risque de produire des images de qualité médiocre augmente considérablement. Il est, par conséquent, impératif que le scanner soit réglé de manière optimale avant la production et que le réglage soit bien maintenu durant l'opération de balayage. Pour cela, nous avons mis au point une procédure de CQ appelée vérification de l'étalonnage du scanner, pour examiner systématiquement le réglage et l'étalonnage de la machine.

Puisque les réglages du matériel du système de balayage ont tendance à ne pas fluctuer fortement, nous avons estimé que des méthodes de contrôle statistique du processus (CSP) seraient appropriées pour surveiller cette partie de l'opération. Il s'agissait essentiellement d'une vérification ponctuelle régulière de l'étalonnage, exécutée quotidiennement avant le début de la production. Le contrôle de l'étalonnage consistait à balayer de nouveau un lot d'essai et à comparer les résultats à ceux obtenus avant l'étalonnage pour le même lot, puis à déterminer les écarts entre les résultats réels et prévus pour calculer les taux d'erreur. Enfin, ces taux d'erreurs ont été portés sur des cartes de contrôle CSP pour déterminer si le processus fonctionnait à un niveau acceptable. Si le lot d'essai était rejeté, l'étape du balayage ne pouvait démarrer qu'après le réétalonnage de la

graveité du problème observé.

Pour le triage, la mise en lots, le découpage et le stockage, la mesure de la qualité choisie était le « pourcentage de questionnaires erronés » (c'est-à-dire conformément aux hypothèses pour une unité d'échantillonnage simple). Pour l'opération de transcription, la probabilité que de multiples erreurs indépendantes surviennent dans un questionnaire était très forte et, par conséquent, la mesure de la qualité choisie était le « nombre de défauts pour cent unités, *NDCU* » (c'est-à-dire, conformément aux hypothèses requises pour une unité d'échantillonnage complexe).

Application du contrôle de la qualité à la saisie des données par RIC : Recensement de l'agriculture du Canada de 2001

WALTER MUDRYK et HANSHENG XIE¹

RÉSUMÉ

La reconnaissance intelligente de caractère (RIC) est une nouvelle technologie de saisie des données d'usage très répandue. Statistique Canada l'a utilisée pour la première fois pour traiter les données du Recensement de l'agriculture du Canada de 2001. Cet exercice a posé de nombreux défis d'ordre tant opérationnel que méthodologique. Le présent article donne un aperçu des outils méthodologiques utilisés pour mettre en place un système de RIC efficace. Puisque le risque d'erreur est élevé aux diverses étapes de l'opération, des méthodes et des procédures d'assurance de la qualité (AQ) et de contrôle de la qualité (CQ) ont été intégrées à celle-ci afin de s'assurer du haut degré d'exactitude des données saisies. L'article décrit ces méthodes d'AQ/CQ ainsi que leur résultat et montre comment ont été réalisées les améliorations de la qualité dans l'opération de saisie des données par RIC. Il souligne aussi les effets positifs de ces procédures sur l'opération de saisie.

MOTS CLÉS : Saisie des données; reconnaissance intelligente de caractère (RIC); contrôle de la qualité; amélioration de la qualité; contrôle statistique du processus.

1. INTRODUCTION

Les données du Recensement de l'agriculture du Canada de 2001 ont été saisies de juillet à novembre 2001 à l'aide d'une nouvelle technologie d'origine assez récente appelée reconnaissance intelligente de caractères (RIC). Cette approche combine la saisie mécanique automatisée fondée sur la reconnaissance optique de caractères, de marques et d'images et la saisie manuelle selon des techniques frontales d'entrée clavier à partir d'une image (« key from image »). La technique frontale de saisie manuelle des données est appliquée uniquement aux zones que le système optique ne peut reconnaître avec un degré suffisant de confiance (c'est-à-dire préalable).

L'adoption de la RIC pour l'opération de saisie des données offrait de nombreux avantages, du point de vue de l'économie des ressources et du gain de productivité. Mais parallèlement, la question de l'exactitude est devenue extrêmement importante dans le cas du traitement d'un grand nombre de documents, car le niveau d'erreur risquait d'être inacceptable à diverses étapes du processus. Quelques auteurs ont étudié la qualité des applications de RIC, dont Karpic (1994) et Pasley (2000). Karpic discute de l'algorithme de codage et des résultats pour l'opération de codage des données du Recensement de 1991 en Croatie et en Bosnie-Herzégovine au moyen de lecteurs optiques intelligents. Pasley fait remarquer que la qualité du balayage optique d'une image dépend de la qualité du document d'origine, de la précision du scanner, de l'expérience de l'opérateur et de la résolution à laquelle est effectué le balayage. Notre objectif étant d'améliorer la qualité des données, nous avons intégré des procédures d'AQ et de CQ dans l'opération de saisie des données du Recensement de

L'agriculture du Canada de 2001 pour assurer un haut degré d'exactitude. Les activités de contrôle de la qualité de l'opération de saisie des données par RIC se sont concentrées sur trois grandes étapes du traitement, à savoir la préparation des documents, l'étalonnage de l'appareil de balayage et la saisie des données figurant sur les questionnaires. Nous avons adopté cette démarche parce que chaque étape dépendait d'une autre et que chacune pouvait donner lieu à des erreurs importantes en bout de ligne. Par conséquent, idéalement, chaque composante devait avoir son propre système de contrôle. Le but du présent article est de décrire la méthodologie et les procédures d'AQ/CQ associées à chacune des grandes étapes de l'opération de saisie des données par RIC, de résumer les résultats obtenus lors de leur application et de montrer comment des améliorations permanentes de la qualité ont été réalisées dans le cadre de l'opération de saisie des données par la RIC.

2. VUE D'ENSEMBLE DU PROGRAMME DE LA QUALITÉ

Pour mieux comprendre la logique qui sous-tend les procédures d'AQ/CQ, il est utile de donner un aperçu de leurs objectifs et méthodes.

2.1 Objectifs

L'objectif global de qualité du projet était de mesurer, contrôler et améliorer en continu la qualité de l'opération complète de saisie des données par RIC. Nous nous sommes proposés d'atteindre cet objectif en mettant en œuvre une série de procédures d'AQ/CQ à chaque étape critique de

- SRINATH, K.P. et CARPENTER, R.M. (1995). Sampling methods for repeated business surveys. Dans *Business Survey Methods*, (Eds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge et P. Kott). New York: John Wiley & Sons, Inc., 171-183.
- THOMPSON, M.E. (1997). *Theory of Sample Surveys*. London: Chapman & Hall.

4. DISCUSSION

Le présent article donne les expressions conditionnelle et inconditionnelle du biais de rétroaction lorsqu'on estime le total au moyen de l'estimateur par extension courant. Nous montrons que le biais de rétroaction peut être important. Même si la proportion d'unités inadmissibles dans la base de sondage est aussi faible que 5 %, l'introduction dans la base de sondage d'information sur ces unités provenant d'enquêtes par sondage peut causer un biais de l'ordre de 2 à 3 %. Toutefois, selon une étude en simulation à petite échelle, si la proportion d'unités inadmissibles est égale ou inférieure à 3 %, la stratégie de rétroaction ne semble pas créer de problème en ce qui concerne le biais ni la variance. Nous dérivons également un estimateur presque sans biais. L'étude en simulation montre qu'en ce qui concerne la variance, cet estimateur se compare favorablement à l'autre stratégie consistant à retenir les unités inadmissibles dans la base de sondage et à permettre qu'elles soient incluses dans les futurs échantillons. Cet estimateur repose sur l'existence d'estimations convergentes des nombres d'unités admissibles et inadmissibles dans la population. Ces estimations peuvent être obtenues d'après un échantillon antérieur grâce à la stratégie sans biais consistant à permettre que les unités pour lesquelles il est établi qu'elles sont disparues soient incluses dans l'échantillon.

Afin de faciliter le développement théorique, nous émettons des hypothèses de simplification. La plus importante est que *toutes* les unités disparues ont été découvertes lors d'une enquête par sondage antérieure et que l'information a été introduite dans la base de sondage. Nous envisageons une base de sondage contenant une zone « blanche », où toutes les unités inadmissibles sont marquées en tant que telles, et une zone « noire », où aucune unité inadmissible n'est touchée. En pratique, il est peu probable que ce genre de situation se produise. Si elle est moyennement grande et qu'elle est utilisée pour de nombreuses enquêtes continues, parmi lesquelles certaines fournissent des renseignements à des degrés divers, la base de sondage deviendra « grise » plutôt que « noire et blanche ». Le biais de rétroaction sera alors moins grave que dans le cas d'une situation « noire et blanche ». Cependant, la quantification du biais d'une base de sondage « raisonnablement grise » dépasse le cadre du présent article. En ce sens, la situation examinée ici représente le pire scénario.

Les auteurs remercient Mark Pont pour les premières discussions fort utiles du sujet qu'ils ont eues avec lui. Ils expriment également leur reconnaissance à un rédacteur adjoint et à deux examinateurs pour leurs commentaires

REMERCIEMENTS

lorsqu'il a participé à l'étude.

National Cancer Institute des États-Unis (CA 57030). Hedlin était employé par l'Université de Southampton

auteurs ont été financés partiellement par l'Office for National Statistics du Royaume-Uni. Les travaux de recherche de Wang ont également été financés par le

BIBLIOGRAPHIE

COLLEDGE, M.J. (1989). Coverage and classification maintenance issues in economic surveys. Dans *Panel Surveys*. (Éds., D. Kasprzyk, G.J. Duncan, G. Kalton et M.P. Singh). New York: John Wiley & Sons, Inc., 80-107.

COLLEDGE, M.J. (1995). Frames and business registers: an overview. Dans *Business Survey Methods*. (Éds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge et P. Kott). New York: John Wiley & Sons, Inc., 21-47.

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. Dans *New Developments in Survey Sampling*. (Éds., N.L. Johnson et H. Smith). New York: John Wiley & Sons, Inc., 629-651.

ERNST, L.R., VALLIANT, R. et CASADY, R.J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics*, 16, 211-228.

HIDRIGLOU, M.A., et LANIEL, N. (2001). Sampling and estimation issues for annual and sub-annual Canadian business surveys. *Revue internationale de Statistique*, 69, 487-504.

HIDRIGLOU, M.A., et SRINATH, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.

LEE, H., RANOURT, C.-E. et SÄRNÄL, (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

LAVALLÉE, P. (1996). Frame update problems with panel surveys. *Proceedings of Statistical Days '96*, Statistical Society of Slovenia, 252-261.

OHLSSON, E. (1995). Coordination of samples using permanent random numbers. Dans *Business Survey Methods*. (Éds., B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge et P. Kott). New York: John Wiley & Sons, Inc., 153-169.

SCHIOPU-KRATINA, I., et SRINATH, K.P. (1991). Renouvellement de l'échantillon et estimation dans l'enquête sur l'emploi, la rémunération et les heures de travail. *Techniques d'enquête*, 17, 89-100.

SRINATH, K.P. (1987). Methodological problems in designing continuous business surveys: some Canadian experiences. *Journal of Official Statistics*, 3, 283-288.

L'examen des probabilités de couverture facilite l'évaluation du biais. Le tableau 2 montre les probabilités de couverture, fondées sur des « intervalles de confiance » symétriques dont la largeur est égale à deux fois l'écart type empirique simulée de part et d'autre de la proportion ponctuelle. Bien que la stratégie 2 produise dans toutes les cellules des probabilités de couverture proches de la valeur cible de 95 %, la stratégie 1 ne donne le même résultat en général que pour la population comptant 3 % d'unités disparues. Sous la stratégie 1, la probabilité de couverture a aussi tendance à être acceptable pour les populations contenant une plus forte proportion d'unités disparues si la moitié de l'échantillon est tirée de la partie de la droite NAP dont les unités disparues ont été éliminées, et l'autre moitié, de la partie de la droite NAP où a été retenue la proportion originale d'unités disparues, puisque le biais négatif provenant de la première moitié de l'échantillon et le biais positif provenant de la deuxième moitié ont tendance à s'annuler.

Tableau 2
Probabilité de couverture, en pourcentage, pour l'estimation du total de Y1. Dans chaque cellule, la première entrée correspond à la probabilité de couverture sous la stratégie 1 et la seconde, à la probabilité de couverture sous la stratégie 2.

<i>P</i> disparue	Moyenne de <i>n_a/n</i>		
	0 %	50 %	100 %
0,03	94,6	94,3	94,8
0,05	93,3	95,2	94,4
0,20	65,9	94,5	93,8
0,50	21,2	95,1	78,4
			94,7
			94,8
			46,1
			94,6
			95,0
			95,1

Tableau 3
Rapport des variances de l'estimateur du total de Y1. Dans chaque cellule, la première entrée correspond à la variance sous la stratégie 2 comparativement à la stratégie 1 et la seconde, à la variance sous la stratégie 3 comparativement à la stratégie 1.

<i>P</i> disparue	Moyenne de <i>n_a/n</i>		
	0 %	50 %	100 %
0,03	1,04	1,00	1,06
0,05	1,08	0,98	1,14
0,20	1,28	0,85	1,27
0,50	1,85	0,52	1,34
			0,58
			0,83
			1,46
			2,24

Tableau 4
Rapport des variances de l'estimateur du total de Y2. Dans chaque cellule, la première entrée correspond à la variance sous la stratégie 2 comparativement à la stratégie 1 et la seconde, à la variance sous la stratégie 3 comparativement à la stratégie 1.

<i>P</i> disparue	Moyenne de <i>n_a/n</i>		
	0 %	50 %	100 %
0,03	1,03	1,00	1,03
0,05	1,06	0,99	1,04
0,20	1,25	0,92	1,15
0,50	1,80	0,65	1,40
			0,50
			1,36

Nous avons calculé la variance des estimations simulées. Les tableaux 3 et 4 donnent les comparaisons de la variance pour Y1 et Y2, respectivement, sous les stratégies 2 et 3 comparativement à la stratégie 1. Comme prévu, dans tous les cas, la stratégie 1 donne une variance plus petite que la stratégie 3. La stratégie 2 donne de bons résultats dans la plupart des cas, mais, étant donné sa plus grande complexité, il semble préférable de choisir la stratégie 1 de rétroaction pour les populations contenant une faible proportion d'unités inadmissibles, disons 3 % ou moins. Par contre, si cette proportion est plus grande que, disons, 5 %, le biais de la stratégie 1 peut donner lieu à des probabilités de couverture médiocres et des estimations incorrectes. La variance de la stratégie 2 n'est pas pire que celle de la stratégie 3; dans la plupart des cas, la stratégie 2 est supérieure. Les rapports des variances non monotones figurant au bas du tableau 3 sont dus à l'estimation de $N_{2,d}$ et de $N_{d,i}$ combinée aux détails particuliers de la simulation.

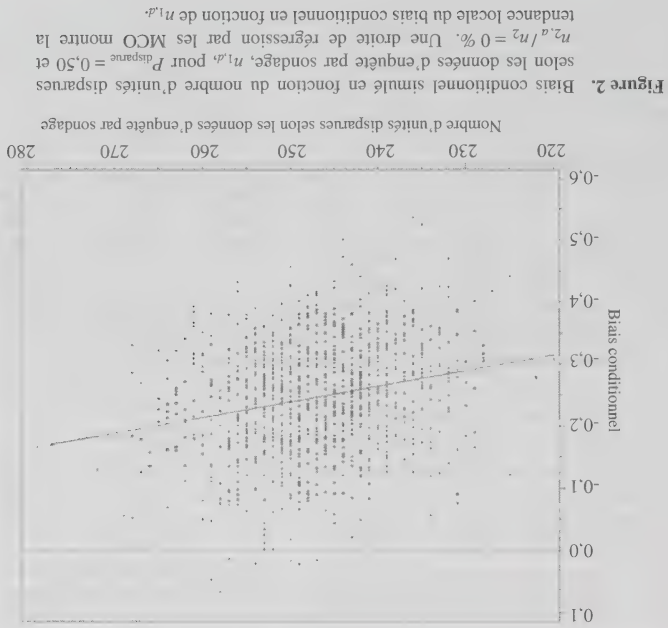
Sommairement, nous avons maintenues fixes les tailles de population et d'échantillons, les variables d'intérêt Y1 et Y2, et les unités particulières qui étaient disparues. Pour douze combinaisons de $P_{disparue}$ et de n_{2a}/n_2 , le réordonnement des unités sur la ligne NAP par simulation de nouveaux NAP a fait varier les facteurs suivants :

- les unités particulières incluses dans s_{1i} , $s_{2courante}$ et s_{2orig} ;
- le nombre d'unités disparues, ainsi que lesquelles de celles-ci sont des unités disparues selon les données d'enquête par sondage;
- les unités particulières qui appartiennent à s_{1i} et U_{2wi} .

Donc, les quantités s_{1a} , N_{2a} et N_2 varient dans les simulations. Il semble pratique de les laisser varier plutôt que de les contrôler dans une expérience comprenant plus de facteurs que $P_{disparue}$ et n_{2a}/n_2 . Donc, les résultats sont inconditionnels, conformément à (8).

Tableau 1
Biais, % du total de Y1. Dans chaque cellule, la première entrée est le biais sous la stratégie 1 et la deuxième, le biais sous la stratégie 2.

$P_{disparue}$	Moyenne de n_a/n			
	0 %	50 %	100 %	
0,03	-1,6	0,4	0,4	1,5
0,05	-2,8	0,0	0,4	2,9
0,20	-10,2	0,2	1,5	12,7
0,50	-24,6	0,2	12,5	49,0



Le tableau 1 montre le biais relatif empirique des stratégies 1 et 2, calculé sous forme de moyenne simple des 1 000 différences entre l'estimation et le paramètre et exprimé en pourcentage du total obtenu par la simulation. La stratégie 3 est sans biais et, par conséquent, n'est pas incluse dans le tableau 1. Le biais empirique de la stratégie 3 qui apparaît néanmoins dans les simulations reflète l'erreur de simulation; il se chiffre, au plus, à 0,5 %. Comme le montre le tableau 1, la stratégie 2 est pratiquement sans biais également. Notons que le biais empirique simulé sous la stratégie 1 correspond à ce que prédit (8) (en tenant compte de l'erreur de simulation). Ce biais est appréciable dans presque tous les cas et, si la proportion d'unités disparues (ou inadmissibles) est élevée, il peut, en effet, être fort important. La figure 2 montre le biais conditionnel sachant n_{1d} pour $P_{disparue} = 0,50$ et $n_{2a}/n_2 = 0 \%$. Notons que le biais donné par (6) est localement bien décrit par la droite de la régression du biais conditionnel sur n_{1d} par les MCO tracée sur la figure. Par exemple, si $n_{1d} = 220$, alors les expressions N_{2a}/N_{2i} et $(n_1 - n_{1d})/(N_1 - n_1)$ sont toutes deux égales à 0,56 et $B = -0,31$.

où s_2 est un échantillon tiré de U_2 . Pour obtenir la stratégie 2, notons que, si nous disposons d'estimations convergentes de $N_{2,d}$ et $N_{2,i}$, nous pouvons les introduire dans (7) ou (8) et obtenir un estimateur ayant de bonnes propriétés

$$\hat{r}_y^* = \hat{r}_y^{*n} (1 + \hat{c})^{-1}, \quad (9)$$

où $\hat{c} = (N_{2,d}/N_{2,i})(n_{2,d}/n_{2,i} - \{n_{1,d}/n_{1,i}\})/\{n_{2,i}/n_{2,i} - \{n_{1,i}/n_{1,i}\}\}$ pour le cas tant conditionnel qu'inconditionnel, puisque, dans (8), le terme $n_{2,d}V(n_{1,d})(n_{2,i}/N_{2,i})^{n_{2,d}})^{-1}$ est presque toujours négligeable. Nous pouvons obtenir les estimations $N_{2,d}$ et $N_{2,i}$ des tailles des domaines $U_{2,d}$ et $U_{2,i}$ à partir d'un échantillon tiré de la population d'enquête originale ou courante. Si nous tirons plus d'un échantillon, nous obtenons des estimations sans biais individuelles de $N_{2,d}$ (ou $N_{2,i}$) que nous pourrions combiner. L'estimateur combiné à variance minimale est la somme des estimateurs pondérés par l'inverse de leur variance. Comme le montre l'argument qui suit, nous ne nous attendons pas à ce que le biais de (9) soit important :

$$E(\hat{r}_y^{*n}) = E[\hat{r}_y^{*n} (1 + \hat{c})^{-1}] \approx E(\hat{r}_y^{*n}) (1 + c)^{-1} = t_y (1 + c)^{-1} = t_y^*.$$

Une autre stratégie, appelée ici stratégie 3, consiste à introduire dans la base de sondage l'information selon laquelle certaines unités sont disparues, mais à garder ces unités dans la base de sondage et à permettre qu'elles soient échantillonnées. L'estimateur résultant est sans biais, mais l'inconvénient de la stratégie est que la précision souffre puisqu'une partie de l'échantillon est perdue sous forme d'unités inadmissibles. L'estimateur de la stratégie 3 est $\hat{r}_y^{*n} = N_1/n_{2,i} \sum_{i \in Y_i} y_i$, où r est un échantillon tiré de la population d'enquête originale U_1 .

3.2 Étude en simulation

Une étude en simulation pourrait fournir des éclaircissements quant à celle des stratégies 1 à 3 qu'il convient d'utiliser. Les mesures qui s'imposent naturellement pour la comparaison des stratégies sont le biais et la variance. Dans le cas des enquêtes auprès des entreprises, les estimations pour des sous-populations (industries) sont souvent plus intéressantes que celles pour l'ensemble de la population. En vue de simuler une sous-population, nous avons créé une base de sondage contenant 1 000 unités pour former la population d'enquête originale. Nous avons associé à chaque unité une valeur d'une variable Y_1 obéissant à une loi gamma. Nous avons utilisé la même loi gamma que celle générée par la population 12 dans Lee, Ramcourt et Samdal (1994, page 236). Le coefficient de variation (écart-type de population divisé par la moyenne) était égal à 0,57. Nous avons créé une autre variable

La procédure décrite au paragraphe qui précède a été répétée 1 000 fois. Autrement dit, pour chaque valeur de $P_{disparue}$ mentionnée plus haut et pour chacun de trois points de départ de s_2 , à définir, nous avons généré 1 000 ensembles de NAP, que nous avons associés aux unités. Nous avons réordonné la base de sondage pour chaque nouvel ensemble de NAP, et tiré trois échantillons pour chaque réordonnement (s_1 , s_2 courante et s_2 orig.). Nous avons choisi deux valeurs de $départ_2$, à savoir 0,0 et 0,7, de sorte que la proportion de s_2 courante comprise dans s_1 soit égale à 100 % ou à 0 %, respectivement. Autrement dit, nous avons fixé $n_{2,d}/n_2$ à 100 % et à 0 %. En outre, nous avons calculé les valeurs appropriées de $départ_2$ pour que $n_{2,i}/n_2$ soit, en moyenne, de 50 % sous chaque $P_{disparue}$ choisie. Ces valeurs sont 0,448, 0,447, 0,438 et 0,4 pour des valeurs de $P_{disparue}$ égales à 0,03, 0,05, 0,2 et 0,5, respectivement.

Nous avons associé un NAP à chaque unité et ordonné les unités le long d'une ligne NAP. Puis, nous avons tiré le premier échantillon, s_1 , en repérant les 500 unités ayant les plus petits NAP. Nous avons marqué toutes les unités disparues dans s_1 comme étant « disparues selon les données d'enquête par sondage ». Donc, s_1 couvre environ la première moitié de la ligne NAP. La base de sondage dont ont été exclues les unités marquées comme étant disparues selon les données d'enquête par sondage constitue la population d'enquête courante. Les estimations de $N_{2,d}$ et $N_{2,i}$ utilisées pour la stratégie 2 sont fondées sur s_1 . Un deuxième échantillon, représenté par s_2 courante, a été tiré en sélectionnant 100 unités à la droite d'un point de départ, $départ_2$, en ne considérant pas les unités disparues selon les données d'enquête par sondage. Nous avons sélectionné un autre échantillon de 100 unités à partir de $départ_2$, mais en permettant cette fois que les unités disparues selon les données d'enquête par sondage soient incluses dans l'échantillon. Donc, cet échantillon a été tiré à partir de U_1 et nous le représentons par s_2 orig. L'échantillon s_2 courante est pertinent pour les stratégies 1 et 2, tandis que s_2 orig sera utilisé pour la stratégie 3.

La procédure décrite au paragraphe qui précède a été répétée 1 000 fois. Autrement dit, pour chaque valeur de $P_{disparue}$ mentionnée plus haut et pour chacun de trois points de départ de s_2 , à définir, nous avons généré 1 000 ensembles de NAP, que nous avons associés aux unités. Nous avons réordonné la base de sondage pour chaque nouvel ensemble de NAP, et tiré trois échantillons pour chaque réordonnement (s_1 , s_2 courante et s_2 orig.). Nous avons choisi deux valeurs de $départ_2$, à savoir 0,0 et 0,7, de sorte que la proportion de s_2 courante comprise dans s_1 soit égale à 100 % ou à 0 %, respectivement. Autrement dit, nous avons fixé $n_{2,d}/n_2$ à 100 % et à 0 %. En outre, nous avons calculé les valeurs appropriées de $départ_2$ pour que $n_{2,i}/n_2$ soit, en moyenne, de 50 % sous chaque $P_{disparue}$ choisie. Ces valeurs sont 0,448, 0,447, 0,438 et 0,4 pour des valeurs de $P_{disparue}$ égales à 0,03, 0,05, 0,2 et 0,5, respectivement.

$$\Pr[k \in s_{2,b} | k \in U_{2,wd}, n_{1,d}] = \frac{N_{2,wd}}{n_{2,b}}. \quad (4)$$

À partir de (4), nous obtenons la valeur prévue conditionnelle de $f_{(s_{2,b})}^{y_{(s_{2,b})}}$, soit

$$E(f_{(s_{2,b})}^{y_{(s_{2,b})}} | n_{1,d}) = E \left[\frac{N_{2,wd}}{n_{2,b}} \sum_{U_2} w_k y_k | n_{1,d} \right] = \frac{N_{2,wd}}{n_{2,b}} \sum_{U_2} w_k y_k.$$

La seconde équation susmentionnée est due au fait que, sachant $n_{1,d}$, les $N_{2,i}$ unités actives dans U_2 sont toutes aussi susceptibles les unes que les autres d'être comprises dans $U_{2,wd}$, qui contient $N_{2,i} - n_{1,d}$ unités actives. Par conséquent, le biais conditionnel de $f_{(s_{2,b})}^{y_{(s_{2,b})}}$ est

$$B(f_{(s_{2,b})}^{y_{(s_{2,b})}}, t_y | n_{1,d}) = \sum_{U_2} \left(\frac{w_k n_{2,b}}{N_{2,i} - n_{1,d}} - 1 \right) y_k. \quad (5)$$

Pour l'estimateur par développement en série $f_{(s_{2,b})}^{y_{(s_{2,b})}}$ avec les poids $w_k = N_2/n_{2,b}$, le biais est

$$B(f_{(s_{2,b})}^{y_{(s_{2,b})}}, t_y | n_{1,d}) = B t_y, \quad (6)$$

où

$$B = \frac{N_2}{N_{2,wd}} \frac{N_{2,i} - n_{1,d}}{N_{2,i} - n_{1,d}} - 1 = \frac{N_{2,wd} N_{2,i}}{N_{2,i}^2} \frac{N_{2,i} - n_{1,d}}{N_{2,i} - n_{1,d}} = \frac{N_{2,wd} N_{2,i}}{N_{2,i}^2} \frac{N_{2,i} - n_{1,d}}{N_{2,i} - n_{1,d}}.$$

Le biais est toujours non positif, puisque $B \leq 0$. Il est facile de voir que B est une fonction croissante de $n_{1,d}$, puisque $N_{2,i} = N_{1,d} - n_{1,d}$, où $N_{1,d}$ est le nombre fixe d'unités disparues dans U_1 . Il est également facile de montrer que l'on atteint la valeur maximale de B quand $s_{1,d}$ comprend toutes les unités disparues comprises dans U_1 , c'est-à-dire quand $n_{1,d} = N_{1,d}$ et, par conséquent, $N_{2,i} = 0$.

2.4 Biais de rétroaction provenant des parties d'échantillon combinées

En combinant (6) et (3), nous obtenons l'expression du biais global de $f_{y_{(s_{2,b})}}^{y_{(s_{2,b})}} = N_2/n_2 \sum_{s_{2,b}} y_k$ suivante

$$B(f_{y_{(s_{2,b})}}^{y_{(s_{2,b})}}, t_y | n_{1,d}) = E(f_{y_{(s_{2,b})}}^{y_{(s_{2,b})}} | n_{1,d}) - t_y = \left(\frac{N_{2,i}}{n_{2,b}} \frac{N_{2,i} - n_{1,d}}{N_{2,i} - n_{1,d}} - \frac{N_{2,i}}{n_{2,b}} \frac{N_{2,wd}}{N_{2,i} - n_{1,d}} \right) t_y = c t_y. \quad (7)$$

Le biais présent dans l'estimateur par extension se résume réellement à ne pas connaître la taille correcte de la population. Dans (3), le biais est dû à la multiplication de la moyenne d'échantillon sur les unités actives par N_2 plutôt que par le nombre inconnu $N_{2,i}$. Le biais provenant des parties d'échantillon $s_{2,d}$ et $s_{2,b}$ sera, en valeur absolue, plus faible que (3) et que (6), respectivement, si certaines unités disparues comprises dans les échantillons tirés de U_1 n'ont pas été reconnues comme étant disparues et, par conséquent, n'ont pas été éliminées. Cela pourrait se produire, par exemple, si la situation des unités non répondantes est difficile à déterminer.

Nous pouvons obtenir directement une analyse non conditionnelle en présence de rétroaction en prenant l'espérance de (7) par rapport à $n_{1,d}$. Donc, incon-

ditionnellement, nous avons

$$E \left(\frac{N_2}{n_2} \sum_{s_{2,b}} y_k \right) - t_y = \left[\frac{N_{1,d} - E(n_{1,d})}{n_{2,b}} \left(\frac{n_2}{N_{2,i}} - \frac{n_{2,b}}{n_1 - E(n_{1,d})} \right) - \frac{n_{2,i}}{n_{2,b}} \frac{N_{2,wd}}{N_{2,i} - n_{1,d}} \right] t_y = c t_y, \quad (8)$$

où $E(n_{1,d}) = n_1 N_{1,d}/N_1$ et $V(n_{1,d}) = n_1 N_{1,d} N_{2,i}/N_1^2$. Lavalée (1996) a adopté une approche intéressante pour résoudre un problème comparable avec des données d'enquête par panel. Dans son article, il traite, entre autres, du problème de la mise à jour de la base de sondage au moyen d'un panel avec renouvellement. Notre approche est différente de la sienne en ce sens que nous considérons les deux probabilités conditionnelles $\Pr[k \in s_{2,b} | n_{1,d}]$ et $\Pr[k \in s_{2,b} | n_{1,d}]$ séparément.

3. TROIS STRATÉGIES SIMPLES ET UNE ÉTUDE EN SIMULATION

3.1 Stratégies en présence de rétroaction

Une stratégie, que nous appelons ici stratégie 1, consiste à procéder à la rétroaction, supprimer l'ensemble $s_{1,d}$ de la base de sondage et accepter le biais de rétroaction. Malheureusement, on connaît rarement la grandeur du biais. Pour la stratégie 1 sous EAS, l'estimateur est $f_{y_{(s_{2,b})}}^{y_{(s_{2,b})}} = N_2/n_2 \sum_{s_{2,b}} y_k$,

parties d'échantillon comme deux échantillons de taille fixe, chacun tiré selon un plan EAS à partir de leur sous-population respective. Nous conditionnons sur $n_{2,a}$ et $n_{2,b}$ dans tout l'exposé sans l'indiquer explicitement dans les formules. Au moyen de la notation ($k \in s_{2,a}$), nous faisons référence à l'événement selon lequel une unité est d'abord incluse dans le ou les échantillons de premier cycle tirés de U_1 , puis dans l'échantillon de deuxième cycle tiré de ce qui reste du ou des échantillons de premier cycle après que les unités disparues aient été éliminées. La notation ($k \in s_{2,b}$) est analogue. Soit $I(k \in s_{2,a}) = 1$ quand l'unité k est incluse dans $s_{2,a}$ et $I(k \in s_{2,a}) = 0$, autrement. Pour calculer le biais global, il est commode d'analyser les biais provenant des parties d'échantillon $s_{2,a}$ et $s_{2,b}$. Nous établissons une expression pour chacun d'eux aux sections 2.2 et 2.3, respectivement, et nous les amalgamons à la section 2.4.

2.2 Biases de rétroaction provenant d'un sous-échantillon tiré de l'échantillon original

Supposons qu'un sous-échantillon $s_{2,a}$ soit tiré de $s_{1,a}$ la partie active du ou des échantillons de premier cycle. Rappelons que $y_k = 0$ si k est une unité disparue et que $U_2 = U_{2,a} \cup U_{2,l}$. Donc, nous avons $\sum_{s_{2,a}} y_k I(k \in s_{2,a}) = \sum_{U_2} y_k I(k \in s_{2,a})$. Supposons que $N_{2,l} > 0$. Alors, nous obtenons que $\Pr[k \in s_{2,a} | n_{1,d}] = n_{2,a} / N_{2,l}$, puisqu'un échantillon de taille $n_{2,a}$ est effectivement sélectionné à partir d'une population de taille $N_{2,l}$ selon le plan EAS (par la voie d'un échantillon EAS tiré de U_1 , suivi d'un échantillon EAS tiré de $U_{2,l}$). Notons

$$B(\hat{t}_{y(s_{2,a})}^{(s_{2,a})}, t_y | n_{1,d} | \{w_k \Pr[k \in s_{2,a} | n_{1,d}] - 1\} y_k) = \sum_{U_{2,l}} \{w_k \Pr[k \in s_{2,a} | n_{1,d}] - 1\} y_k \\ = \sum_{U_{2,l}} \left(\frac{N_{2,l}}{w_k n_{2,a}} - 1 \right) y_k. \quad (2)$$

Pour la partie d'échantillon $s_{2,a}$, l'estimateur par extension naïf qui ne tient pas compte du biais de rétroaction aurait les poids $w_k = N_2 / n_{2,a}$. D'après (2), le biais de l'estimateur $\hat{t}_{y(s_{2,a})}^{(s_{2,a})} = N_2 / n_{2,a} \sum_{s_{2,a}} y_k$ est

$$B(\hat{t}_{y(s_{2,a})}^{(s_{2,a})}, t_y | n_{1,d}) = \frac{N_{2,l}}{N_{2,d}} t_y. \quad (3)$$

2.3 Biases de rétroaction provenant d'un nouvel échantillon tiré à partir de la population d'enquête courante

Maintenant, nous calculons le biais dû à la partie d'échantillon $s_{2,b}$ de taille $n_{2,b}$ tiré à partir de U_2 par la voie de $U_{2,w}$ (voir la figure 1). Commençons par noter que

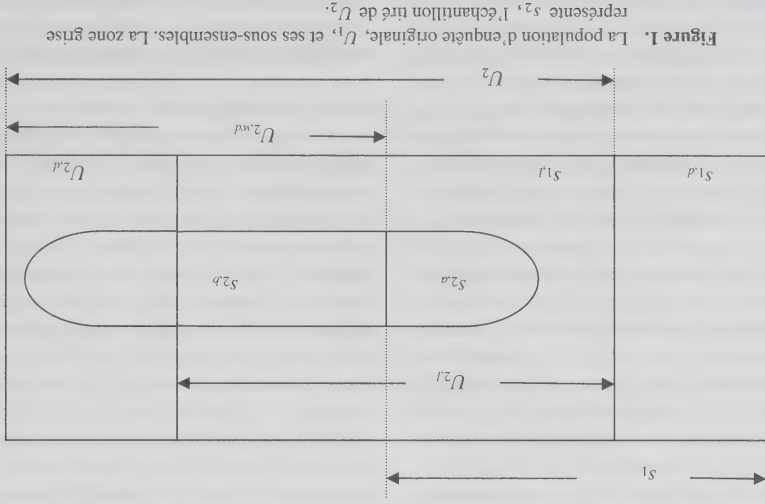


Figure 1. La population d'enquête originale, U_1 , et ses sous-ensembles. La zone grise représente s_2 , l'échantillon tiré de U_2 .

2.1 Introduction et notation

échantillonage, mais pour lesquelles l'indépendance des mécanismes de détection et d'échantillonnage ne peut être confirmée sont appelées *échantillons selon les données* (*d'enquête par sondage*). Dans notre cadre de travail, il s'agit des unités disparues décelées dans les échantillons tirés à partir de U_1 . Soit $s_{1,d}$ l'ensemble de ces unités et nous avons la relation $U_1 = U_2 \cup s_{1,d}$. La figure 1 montre les ensembles et les relations entre ces derniers. Soit N et n , avec un indice inférieur approprié, la taille de la population et du ou des échantillons correspondants, respectivement. Alors, $N_1 = N_2 + n_{1,d}$ et $N_2 = N_{2,d} + N_{2,d}$. Au moment où les échantillons sont tirés à partir de U_2 , N_2 et $n_{1,d}$ sont des nombres connus, tandis que $N_{2,d}$ et $N_{2,d}$ sont inconnus. En outre, $n_{1,d}$, $N_{2,d}$ et N_2 pourraient être considérés comme aléatoires selon les résultats de la rétroaction, tandis que $N_{2,d}$ est fixe. Suivant les principes énoncés dans Durbin (1969) et, plus récemment, dans Thompson (1997), dans nombre de situations, nous pourrions conditionner sur $n_{1,d}$. Par exemple, si il est constaté que $n_{1,d} = 0$, alors il ne semble pas approprié d'inclure dans l'inférence la possibilité que $n_{1,d}$ pourrait avoir été grand. Cependant, à l'étape de la planification de l'enquête, une analyse inconditionnelle serait préférable pour analyser le développement du biais de rétroaction au cours d'une série de cycles d'une enquête par panel. Nous donnons aussi une expression pour le biais de rétroaction inconditionnel.

Soit $s_{1,l}$ la partie active de $s_{1,l}$, c'est-à-dire la partie de U_2 couverte par le ou les échantillons tirés antérieurement de U_1 ; voir la figure 1. De toute évidence, $s_{1,l}$ est un ensemble aléatoire et nous avons $s_{1,l} \subset U_{2,l}$. Soit $U_{2,w,d}$ la partie non échantillonnée de U_2 (« wd » pour « with dead units », y compris les unités disponibles). Il s'agit également d'un ensemble aléatoire qui englobe l'intervalle de $U_{2,d}$ et une partie de $U_{2,l}$. Nous avons $U_2 = U_{2,w,d} \cup s_{1,l}$.

Soit s_2 un EAS tiré de U_2 . Les estimateurs fondés sur s_2 seront entachés d'un biais de rétroaction à moins que l'on dispose d'information spéciale, telle que la connaissance de $N_{2,d}$, ce qui n'est habituellement pas le cas. Pour dériver une expression pour le biais de rétroaction, nous commençons par obtenir les probabilités d'inclusion. Pour cela, il est utile de considérer les deux parties de l'échantillon s_2 séparément : la partie $s_{2,d}$ de taille $n_{2,d}$ tirée de $s_{1,l}$ par échantillonnage NAP ou par une technique d'échantillonnage par panel, et la partie restante $s_{2,p}$ tirée à partir de $U_{2,w,d}$. Si l'échantillonnage est fait selon une méthode par panel, les parties d'échantillon $s_{2,d}$ et $s_{2,p}$ correspondent à l'ancien et au nouveau groupes de renouvellement, respectivement. Si l'échantillon est tiré par échantillonnage NAP, $s_{2,d}$ et $s_{2,p}$ comprennent les unités dont le NAP tombait dans s_1 ou ne tombait pas dans s_1 , respectivement. Que l'échantillon soit tiré par échantillonnage NAP ou par échantillonnage par panel, nous pouvons considérer l'

les disparitions au premier cycle, l'échantillon de deuxième cycle contient effectivement moins d'information sur la proportion d'unités actives dans la population. Nous montrons que l'estimation du nombre d'unités actives dans la population peut être utilisée pour produire une estimation approximativement sans biais du total de population.

Une recommandation prudente serait de n'utiliser aucune autre information sur les unités disparues tirées des enquêtes par sondage que celle provenant des strates entièrement dénombrées en vue de mettre à jour la base de sondage quand les échantillons sont coordonnés au cours du temps (voir Ohlsson 1995, page 168, et Colledge 1989, page 103). Toutefois, il semble qu'interdire la réaction équivaldrait à ne pas se donner le droit d'utiliser toute l'information disponible. Nous obtenons une expression pour le biais dû à la réaction et montrons que celui-ci peut être estimé et utilisé pour corriger les estimateurs conventionnels. Schopru-Kraïna et Srinath (1991) corrigent les poids de sondage pour contrecarrer une proportion trop faible prévue d'unités disparues dans l'échantillon de renouvellement de l'Enquête sur l'emploi, la rémunération et les heures réalisées par Statistique Canada. Hidiroglou et Laniet (2001) discutent brièvement de la question de la réaction. Colledge (1995) discute de façon générale des questions liées à la base de sondage, et Colledge (1989), Hidiroglou et Srinath (1993), Srinath et Carpenter (1995), ainsi que Hidiroglou et Laniet (2001) donnent un aperçu des questions que soulèvent les enquêtes continues auprès des entreprises.

Au lieu des termes admissible et inadmissible, nous utilisons les termes plus émotifs disparue et active, mais notre raisonnement couvre toutes les formes d'inadmissibilité. Nous limitons la discussion à l'estimation du total

$$(1) \quad \sum_{U \in \mathcal{U}} y_k = t_k$$

d'une variable d'intérêt $y' = (y_1, y_2, \dots, y_N)$ dans une population U dont les étiquettes d'unité sont $\{1, 2, \dots, N\}$.

Lors de l'observation des unités échantillonnées, nous supposons que *toutes* les unités disparues incluses dans l'échantillon sont classées comme étant disparues et que la base de sondage est mise à jour au moyen de cette information. Cet exercice peut être difficile en pratique. Néanmoins, dans le cadre de certaines enquêtes, l'admissibilité de toutes les unités non répondantes peut être déterminée correctement.

À la section 2, nous introduisons la notation et les concepts nécessaires, et donnons les expressions pour l'estimation du biais de rétroaction lors de l'estimation d'un total. À la section 3, nous discutons de trois stratégies qui pourraient être utilisées en présence de rétroaction et les comparons au moyen d'une étude en simulation. Nous concluons l'article par une discussion à la section 4.

En principe, deux grandes catégories de données sont sous contrôle qu'à l'obtention de plans de sondage différents avantage de la façon dont le chevauchement des échantillons NAF et l'échantillonnage par panel tient dactenus dans l'échantillon. La différence entre l'échant-nouveau lors du deuxième cycle, tandis que les autres sont utilisés pour tenir à jour une base de sondage, à savoir des données administratives et des données d'enquête. Divers organismes administratifs envoient régulièrement à l'ONS des bandes magnétiques contenant de l'information sur, par exemple, les créations et les disparitions d'entreprises. Bien que ces bandes magnétiques soient transmises à l'ONS très fréquemment, la distribution du temps nécessaire pour qu'une nouvelle unité ou la modification d'une ancienne soit enregistrée dans la base de sondage est fortement asymétrique. Cette situation est due, en partie, aux procédures de tenue à jour de la base de sondage, par exemple pour éviter les enregistrements en double. En outre, très souvent, il existe un intervalle important entre la cessation réelle et la cessation officielle des activités d'une entreprise. Par conséquent, la plupart des programmes d'enquêtes auprès des entreprises de l'ONS partagent avec d'autres l'information sur les disparitions d'entreprises qu'ils obtiennent d'après leurs échantillons afin d'accélérer le processus d'information. Nous examinons les effets de l'emploi de données d'enquête par sondage pour mettre à jour une base de sondage utilisée pour des enquêtes répétées. Ce moyen est celui adopté en principe par l'ONS, Statistique Suède et d'autres instituts nationaux de la

statistique pour traiter l'information sur les unités disparues dans les enquêtes auprès des entreprises.

Il semble naturel que cette nouvelle information soit mise à la disposition d'autres programmes d'enquêtes par sondage qui, autrement, risqueraient d'inclure les unités disparues dans leurs échantillons et, par conséquent, d'obtenir des résultats moins précis. Cependant, comme le souligne, entre autres, Srinath (1987), ce genre de procédure peut introduire un biais. Ce dernier, que nous appelons *biais de rétroaction*, se produit quand le mécanisme d'échantillonnage n'est pas indépendant des procédures de rétroaction de l'information dans la base de sondage. Ainsi, considérons une situation où toutes les unités disparues sont déjettées et supprimées au moment du premier cycle d'une enquête par panel. Si aucune autre disparition n'a lieu jusqu'au moment de l'observation des unités du panel lors du deuxième cycle, l'échantillon de deuxième cycle contient uniquement des unités actives. Si l'on ne connaît pas le nombre total d'unités actives dans la population au moment du deuxième cycle, on ne peut établir un estimateur sans biais du total. Bien qu'on ait recueilli un plus grand nombre de renseignements sur la population lorsqu'on a enregistré

Introduction dans la base de sondage de l'information sur l'inadmissibilité provenant des enquêtes

DAN HEDLIN et SUOJIN WANG¹

RÉSUMÉ

Il arrive souvent de découvrir à l'étape de la collecte des données d'une enquête que certaines unités de l'échantillon ne satisfont pas aux critères d'admissibilité, alors que l'information enregistrée dans la base de sondage indique le contraire. Par exemple, dans le cas des enquêtes auprès des entreprises, il est fréquent qu'une proportion non négligeable d'unités échantillonnées aient mis fin à leurs activités commerciales depuis la dernière mise à jour de la base de sondage. Cette information peut être enregistrée dans la base de sondage et utilisée lors des enquêtes subséquentes, de façon à rendre les unités inadmissibles comprises dans l'échantillon (ou l'ensemble d'échantillons) lors du premier de deux cycles d'une enquête, nous supposons que toutes les unités inadmissibles comprises dans l'échantillon (ou l'ensemble d'échantillons) sont décélées et exclues de la base de sondage. Lors du deuxième cycle, nous observons de nouveau un sous-échantillon de la partie admissible. Le sous-échantillon peut être agrandi au moyen d'un nouvel échantillon qui contiendra à la fois des unités admissibles et inadmissibles. Nous étudions l'effet que peut avoir sur l'estimation le processus d'introduction de l'information sur l'inadmissibilité dans la base de sondage et nous établissons une expression du biais qui peut résulter de cette rétroaction. Nous nous concentrons sur l'estimation du total au moyen de l'estimateur par extension courant. Nous obtenons un estimateur presque sans biais en présence de rétroaction. Cet estimateur dépend de la disponibilité d'estimations convergentes des nombres d'unités admissibles et inadmissibles dans la population.

MOTS CLÉS : Unité disparue; biais de rétroaction; suréchantillonnage; échantillonnage fondé sur des nombres aléatoires permanents; enquête par panel; échantillons coordonnés.

1. INTRODUCTION

Pour faciliter l'estimation des changements, on veille habituellement à ce que les échantillons consécutifs d'une enquête répétée se chevauchent. Quant on tire les échantillons de plusieurs enquêtes à partir d'une même base de sondage, il est généralement souhaitable de répartir le fardeau de réponse en s'assurant que les échantillons sélectionnés pour les diverses enquêtes ne se chevauchent pas plus qu'il ne faut. Cette précaution est particulièrement importante si la base de sondage est de grandeur moyenne et utilisée pour de nombreuses enquêtes continues, problème qui se pose à nombre d'instituts nationaux de la statistique lorsqu'ils réalisent des enquêtes auprès des entreprises. L'échantillonnage aléatoire simple stratifié est un plan de sondage utilisé très fréquemment pour ce genre d'enquêtes. Étant donné la distribution asymétrique des entreprises, la fraction d'échantillonnage doit être grande pour de nombreuses strates, ce qui aggrave le fardeau de réponse des moyennes et des grandes entreprises. Dans le domaine de la statistique officielle des entreprises, les questions de l'estimation du changement et du fardeau de réponse sont toutes deux d'une grande importance. Par conséquent, les organismes statistiques ont établi des systèmes d'échantillonnage qui leur permettent de coordonner les échantillons, positivement ou négativement (c'est-à-dire créer des

chevauchements ou s'assurer que le chevauchement soit

faible).

Ainsi, l'Office for National Statistics (ONS) du Royaume-Uni utilise la technique des nombres aléatoires permanents (NAP), qui est une méthode très répandue pour le tirage d'échantillons à partir de listes. Un NAP provenant de la distribution uniforme sur $[0, 1]$ est associé à chaque unité de la base de sondage indépendamment les uns des autres et indépendamment des étiquettes d'unité et de toute variable associée aux unités. Chaque unité garde pendant toute son existence le NAP qui lui est attribué. Les unités peuvent être ordonnées selon une ligne commençant à 0 et se terminant à 1 à laquelle est donné le nom de *ligne NAP*. Pour tirer un échantillon aléatoire simple sans remise, auquel nous donnons le nom de *EAS*, de taille n pré-déterminée, nous choisissons un point (aléatoirement ou intentionnellement) sur la ligne NAP et nous incluons les n unités situées, disons, à la droite de ce point dans l'échantillon. Deux EAS sont entièrement coordonnés s'ils sont sélectionnés à partir du même intervalle. Pour une vue d'ensemble de la méthode et d'autres précisions, consulter Ohlsson (1993) et Emsw, Valliant et Casady (2000).

Les échantillons utilisés pour des enquêtes répétées peuvent également être sélectionnés selon une technique de panel, selon laquelle un ensemble de groupes de renouvellement sont sélectionnés au moment du premier cycle

- DALENUS, T., et HODGES, J.L. (1957). The choice of stratification points. *Skandinavisk Aktuarietidskrift*, 198-203.
- DALENUS, T., et HODGES, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 88-101.
- DORFMAN, A.H., et VALLIANT, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.
- ECKMAN, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 219-229.
- FALK, E., ROTZ, W., et YOUNG, L.L.P. (2003). Stratified sampling for sales and use tax highly skewed data-determination of the certain stratum cut-off amount. *Proceedings of the Section on Statistical Computing*, American Statistical Association, 66-72.
- HEDLIN, D. (2000). A procedure for stratification by an extended ekman rule. *Journal of Official Statistics*, 16, 15-29.
- HORGAN, J.M. (2003). A list sequential sampling scheme with applications in financial auditing. *IMA Journal of Management Mathematics*, 14, 1-18.
- LAVALLE, P., et HIDIROGLOU, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.
- NEYMAN, J. (1934). On the two different aspects of the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.
- NICOLINI, G. (2001). A method to define strata boundaries. Working Paper 01-2001-marzo, Dipartimento di Economia Politica e Aziendale, Università degli Studi di Milano.
- RIVEST, L.-P. (2002). Une généralisation de l'algorithme de Lavalée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214.
- DALENUS, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 203-213.
- COCHRAN, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 345-358.

BIBLIOGRAPHIE

Ces travaux ont été financés par une bourse de l'Irish Research Council for Science, Engineering and Technology. Nous remercions les examinateurs de leurs commentaires et nous ont permis d'améliorer considérablement le document original.

REMERCIEMENTS

L'utilisation de la racine carrée des fréquences, dont l'utilisation est courante, en se servant de quatre populations réelles positivement asymétriques subdivisées en trois, quatre ou cinq strates montrent une augmentation considérable de la précision de l'estimateur de la moyenne, l'augmentation la plus importante survenant lorsque le nombre de strates est égal à cinq. Des comparaisons à la méthode de Lavalée-Hidiroglou indiquent que, dans la plupart des cas, cette méthode nécessite une plus grande taille d'échantillon pour obtenir la même précision que celle donnée par la méthode géométrique; l'augmentation la plus importante de la taille d'échantillon est celle requise pour le nombre le plus grand de strates. Une limite du nouvel algorithme comparativement à la méthode de construction de strates de Lavalée-Hidiroglou est qu'il ne détermine pas de strate supérieure à tirage complet.

Tableau 7 (suite)
Bornes et taille d'échantillon requises avec la méthode de Lavallée-Hidiroglou pour obtenir les mêmes CV qu'avec la méthode géométrique quand $n = 100$

Population		n	CV	4 States			
1	1	113	0,0430	k_h	442	1 828	8 411
				N_h	2 086	915	327
				n_h	16	21	35
				CV_h	0,64	0,41	0,45
2	2	117	0,0194	k_h	19	37	95
				N_h	393	420	176
				n_h	21	34	49
				CV_h	0,19	0,16	0,28
3	3	103	0,0214	k_h	740	1 505	3 819
				N_h	256	234	118
				n_h	9	10	15
				CV_h	0,32	0,18	0,25
4	4	93	0,0142	k_h	117	188	359
				N_h	111	112	74
				n_h	7	9	17
				CV_h	0,14	0,12	0,19
5 States							
1	1	90	0,0360	k_h	342	1 153	3 431
				N_h	1 846	993	357
				n_h	12	14	17
				CV_h	0,58	0,34	0,31
2	2	136	0,0144	k_h	14	21	35
				N_h	189	270	336
				n_h	4	7	16
				CV_h	0,12	0,10	0,12
3	3	105	0,0184	k_h	512	869	1 577
				N_h	133	180	185
				n_h	4	5	10
				CV_h	0,27	0,15	0,16
4	4	99	0,0119	k_h	99	130	189
				N_h	70	68	85
				n_h	4	4	8
				CV_h	0,10	0,08	0,10
				N_h	63	63	71
				n_h	63	63	20
				CV_h	0,33	0,18	0,33

Tableau 6
Efficacités de Cum \sqrt{f} comparativement à la méthode géométrique

Strates	Population			
	1	2	3	4
3	0,97	0,99	0,79	1,16
4	1,23	1,19	1,16	1,04
5	0,94	1,69	1,33	1,17

Nous voyons aussi, en examinant le tableau 6, que, bien que l'efficacité relative soit inférieure à 1 dans quatre cas, pour tous sauf un, elle est supérieure à 0,9. L'exception est la population 3 avec $L = 3$, soit le nombre le plus faible de strates, l'efficacité relative étant de 0,79 dans ce cas.

3.3 Comparaison avec l'algorithme de Lavallée et Hidiroglou

Dans le cas de l'algorithme de Lavallée-Hidiroglou, les bornes optimales k_1, k_2, \dots, k_{L-1} sont choisies de façon à réduire au minimum la taille d'échantillon n pour un niveau donné de précision. L'exigence concernant la précision est habituellement énoncée en spécifiant que le coefficient de variation doit être égal à un niveau donné compris entre 1 % et 10 %. L'obtention du n minimal est un processus itératif et nous avons obtenu le code SAS pour l'exécuter sur Internet à <http://www.ulval.ca/pages/lpr/>.

Tableau 7
Bornes et taille d'échantillon requises avec la méthode de Lavallée-Hidiroglou pour obtenir les mêmes CV qu'avec la méthode géométrique quand $n = 100$

Population		n		CV		3 Strates	
						1	2
1	121	0,0600	k_h	1 248	8 676	38	38
2	123	0,0270	k_h	795	202	41	41
3	107	0,0317	k_h	1 398	4 197	0,17	0,37
4	100	0,0184	k_h	212	85	60	60

Pour comparer les performances de la nouvelle méthode à celles de Lavallée-Hidiroglou, nous utilisons les CV produits par l'algorithme géométrique présentés au tableaux 2, 3 et 4 comme données d'entrée pour l'algorithme de Lavallée-Hidiroglou et nous calculons les tailles d'échantillon nécessaires pour obtenir la même précision que celle de la méthode géométrique avec $n = 100$. Les résultats sont présentés au tableau 7.

La première chose que nous notons en examinant le tableau 7 est que la taille d'échantillon nécessaire pour obtenir la même précision avec l'algorithme de Lavallée-Hidiroglou qu'avec la méthode géométrique est supérieure à 100 dans tous les cas sauf quatre. Pour la population 2 avec cinq strates, il faut augmenter la taille d'échantillon de 36 % de sorte que $n = 136$ pour obtenir la même précision que celle donnée par la méthode géométrique avec $n = 100$. Avec trois et quatre strates, des tailles d'échantillon de $n = 121$ et 113 sont nécessaires pour la population 1 et des tailles d'échantillon de $n = 123$ et $n = 117$, pour la population 2, afin d'obtenir la même précision que celle de la méthode géométrique. Quand la taille d'échantillon devient inférieure à $n = 100$, l'écart n est plus aussi important. Pour la population 4, avec quatre et cinq strates, $n = 93$ et $n = 99$ respectivement, et pour la population 1 avec cinq strates, une taille d'échantillon de $n = 90$ suffit pour obtenir la même précision avec l'algorithme de Lavallée-Hidiroglou qu'avec la méthode géométrique.

Tableau 3 (suite)
Méthode géométrique c. Cum \sqrt{f} : bornes de stratification pour $L = 4$ et $n = 100$

Population	Méthode de stratification	CV	1	2	3	4
4	Géométrique	0.0142	k_h 134	261	504	29
			N_h 156	109	63	29
			n_h 20	23	29	28
	Cum \sqrt{f}	0.0143	CV_h 162	255	488	35
			k_h 207	58	57	35
			N_h 33	9	23	0.24
			CV_h 0.18	0.19	0.18	

Tableau 4
Méthode géométrique c. Cum \sqrt{f} : bornes de stratification pour $L = 5$ et $n = 100$

Population	Méthode de stratification	CV	1	2	3	4	5
1	Géométrique	0.0360	k_h 147	549	2 037	7 552	51
			N_h 1 054	1 267	732	265	24
			n_h 2	14	27	33	51
	Cum \sqrt{f}	0.0349	CV_h 1 644	1 010	332	249	134
			k_h 279	838	1 677	4 193	55
			N_h 364	418	130	87	39
			n_h 18	28	17	20	17
	Géométrique	0.0144	CV_h 0.52	0.30	0.20	0.25	0.57
			k_h 17	32	59	108	39
			N_h 729	92	89	104	40
			n_h 58	4	7	16	15
	Cum \sqrt{f}	0.0186	CV_h 0.18	0.14	0.15	0.16	0.15
			k_h 28	38	57	104	40
			N_h 364	418	130	87	39
			n_h 18	28	17	20	17
	Géométrique	0.0184	CV_h 0.28	0.08	0.11	0.16	0.16
			k_h 433	941	2 043	4 434	56
			N_h 100	255	1 989	74	35
			n_h 2	16	27	20	35
	Cum \sqrt{f}	0.0212	CV_h 1 179	1 669	3 139	6 079	15
			k_h 50	3	17	15	15
			N_h 443	941	2 043	4 434	56
			n_h 2	16	27	20	35
	Géométrique	0.0110	CV_h 0.40	0.09	0.20	0.19	0.13
			k_h 118	200	339	576	24
			N_h 114	116	64	39	24
			n_h 12	20	24	18	24
	Cum \sqrt{f}	0.0119	CV_h 0.14	0.14	0.17	0.12	0.16
			k_h 162	255	395	627	19
			N_h 207	58	37	36	19
			n_h 44	11	10	19	16
			CV_h 0.23	0.11	0.10	0.13	0.11

Tableau 5
Variabilité des CV_h pour les méthodes géométrique et Cum \sqrt{f}

Strates	1	2	3	4
3	0.035	0.050	0.036	0.038
	Cum \sqrt{f}	0.181	0.045	0.035
4	0.027	0.010	0.006	0.008
	Cum \sqrt{f}	0.276	0.062	0.059
5	0.018	0.015	0.013	0.020
	Cum \sqrt{f}	0.166	0.076	0.054

Tableau 2
Méthode géométrique c. Cum \sqrt{f} : bornes de stratification pour $L = 3$ et $n = 100$

Population	Méthode de stratification	CV	1	2	3	Strate
I	Géométrique	0,0600	k_h	354	3 152	189
			N_h	2 334	1 288	45
			n_h	9	46	189
			CV_h	0,71	0,68	0,64
			k_h	558	2 236	295
			N_h	2 339	735	64
2	Géométrique	0,0270	k_h	26	72	94
			N_h	701	243	35
			n_h	36	29	94
			CV_h	0,28	0,23	0,33
			k_h	28	66	101
			N_h	729	208	101
3	Géométrique	0,0317	k_h	9	38	103
			N_h	253	321	103
			n_h	9	38	53
			CV_h	0,32	0,37	0,39
			k_h	1 179	3 629	69
			N_h	456	152	28
4	Géométrique	0,0184	k_h	168	405	53
			N_h	211	93	53
			n_h	27	27	46
			CV_h	0,23	0,24	0,30
			k_h	162	441	43
			N_h	207	107	36
Cum \sqrt{f}		0,0198	CV_h	25	39	0,27
			n_h	23	39	0,27
			k_h	162	441	43
			N_h	207	107	36
			n_h	27	27	46
			CV_h	0,23	0,24	0,30

Tableau 3

Méthode géométrique c. Cum \sqrt{f} : bornes de stratification pour $L = 4$ et $n = 100$

Population	Méthode de stratification	CV	1	2	3	Strate
I	Géométrique	0,0430	k_h	205	1 037	88
			N_h	1 416	1 382	32
			n_h	6	22	88
			CV_h	0,45	0,48	0,50
			k_h	558	1 117	222
			N_h	2 339	483	62
2	Géométrique	0,0194	k_h	20	43	200
			N_h	459	398	51
			n_h	22	31	22
			CV_h	0,22	0,20	0,22
			k_h	19	38	62
			N_h	393	428	29
Cum \sqrt{f}		0,0213	CV_h	15	26	62
			n_h	15	26	29
			k_h	138	343	69
			N_h	526	1 386	42
			n_h	5	27	48
			CV_h	0,27	0,26	0,27
3	Géométrique	0,0214	k_h	20	43	200
			N_h	459	398	51
			n_h	22	31	22
			CV_h	0,22	0,20	0,22
			k_h	19	38	62
			N_h	393	428	29
Cum \sqrt{f}		0,0230	CV_h	15	26	62
			n_h	15	26	29
			k_h	138	343	69
			N_h	526	1 386	42
			n_h	5	27	48
			CV_h	0,27	0,26	0,27

Tableau 1
Statistiques sommaires pour les populations réelles

Population	N	Intervalle	Asymétrie	Moyenne	Variance
1	3 369	40–28 000	6,44	838,64	3 511 827
2	1 038	10–200	2,88	32,57	924
3	677	200–10 000	2,46	1 563,00	3 236 602
4	357	70–1 000	2,08	225,62	36 274

3.2 Comparaison à la méthode de la fonction cumulative de la racine carrée des fréquence

Nous commençons par comparer les performances du nouvel algorithme à cum \sqrt{f} en divisant les populations résumées au tableau 1 en $L = 3$, 4 et 5 strates suivant les deux méthodes pour déterminer les points de coupure. Les résultats sont présentés aux tableaux 2, 3 et 4.

Un examen rapide des coefficients de variation des

tableaux 2, 3 et 4 donne à penser que, dans la plupart des cas, la méthode géométrique donne de meilleurs résultats que cum \sqrt{f} pour l'obtention de CV_h de strate quasi

égaux. Par exemple, pour la population 1, dont le coefficient d'asymétrie est le plus élevé, les CV_h diffèrent consid-

ablement les uns des autres si l'on utilise cum \sqrt{f} pour déterminer les coupures, tandis que la méthode géométrique

semble produire des CV_h quasi égaux dans tous les cas où le nombre de strates est égal à 3, 4 ou 5, les meilleurs

résultats étant obtenus pour $L = 5$. Pour les trois autres populations, les CV_h ne diffèrent pas autant si l'on utilise

cum \sqrt{f} , mais ils continuent de paraître plus variables que ceux obtenus par la méthode géométrique.

Les CV_h obtenus par la méthode géométrique sont plus homogènes si $L = 4$ ou 5 que si $L = 3$, ce qui n'est pas

renforcée lorsque le nombre de strates augmente. Une analyse plus détaillée de la variabilité des CV_h entre

strates est donnée au tableau 5, où l'écart-type des CV_h est calculé pour chaque plan d'expérience.

En examinant le tableau 5, nous constatons qu'à deux exceptions près, les écarts-types des CV_h sont nettement

plus faibles si les strates sont construites par la méthode géométrique que si elles le sont par la méthode cum \sqrt{f} .

Dans les deux cas où cette dernière produit un écart-type plus faible que la méthode géométrique, les écarts ne sont

pas importants et se produisent pour le nombre de strates le plus faible, $L = 3$, dans les populations 2 et 4. Par con-

séquent, nous pouvons conclure que le nouvel algorithme permet de délimiter les strates de telle façon que les CV_h soient quasiment égaux.

Il nous reste à déterminer si les coupures géométriques produisent des estimations plus efficaces que cum \sqrt{f} .

Pour cela, nous comparons les deux méthodes en ce qui concerne l'efficacité relative ou le ratio des variances obtenu pour $n = 100$ réparti de façon optimale entre les strates en utilisant la *méthode de répartition de Neyman* (Neyman 1934) :

$$n_h = \left(\frac{N_h S_{xh}}{N_h S_{xh} + \sum_{i=1}^L N_i S_{xi}} \right) n. \quad (12)$$

L'efficacité relative est définie comme étant

$$ef f_{\text{geom}} = \frac{V_{\text{cum}}(\bar{x}_{st})}{V_{\text{geom}}(\bar{x}_{st})}, \quad (13)$$

où $V_{\text{cum}}(\bar{x}_{st})$ et $V_{\text{geom}}(\bar{x}_{st})$ sont les variances de la moyenne pour la méthode de la fonction cumulative de la racine carrée des fréquences et la méthode géométrique, respectivement, avec $n = 100$ et n_h réparti conformément à (12) pour chaque méthode de stratification. Lors de la planification des tailles d'échantillon, les efficacités relatives peuvent être interprétées comme étant l'augmentation ou la diminution proportionnelle de la taille d'échantillon en utilisant cum \sqrt{f} pour obtenir la même précision que celle donnée par la méthode géométrique avec $n = 100$.

Les calculs de la variance sont fondés sur la variable auxiliaire X et, puisque nous supposons qu'elle est fortement corrélée à la variable étudiée inconnue Y , nous pouvons supposer que l'efficacité relative $ef f$, donnée par (13), est une approximation raisonnable de l'efficacité relative de Y .

Le tableau 6 donne le ratio des variances lorsque le nombre de strates est $L = 3$, 4 ou 5.

Le tableau 6 montre que, si cette nouvelle méthode de construction de strates n'est pas systématiquement plus efficace de celle de la fonction cumulative de la racine

carrée des fréquences, lorsqu'elle l'est, elle l'est nettement plus, et lorsqu'elle ne l'est pas, elle n'est que marginalement

moins bonne. Par exemple, nous observons des gains d'efficacité importants quand $L = 5$ pour les populations 2, 3 et 4 : ici, les efficacités relatives sont 1,69, 1,33 et 1,17, respectivement, ce qui indique qu'il est nécessaire d'utiliser

des échantillons de taille $n = 169$, 133 et 117 avec cum \sqrt{f} pour obtenir la même précision d'échantillon que pour la méthode géométrique avec $n = 100$.

populations asymétriques évoquées par Cochran (1961) pour illustrer l'efficacité de la construction de strates par la méthode de la fonction cumulative de la racine carrée des fréquences, c'est-à-dire :

- la population, exprimée en milliers, des villes américaines (population 2);
 - le nombre d'étudiants dans les programmes d'études de quatre ans des collègues américains (population 3);
 - les ressources, exprimées en millions de dollars, d'une grande banque commerciale américaine (population 4).
- Dans son article, Cochran décrit cinq autres populations auxquelles notre algorithme ne peut s'appliquer. Dans trois cas, la variable est une proportion, à savoir des prêts agricoles, des prêts immobiliers et des prêts indépendants

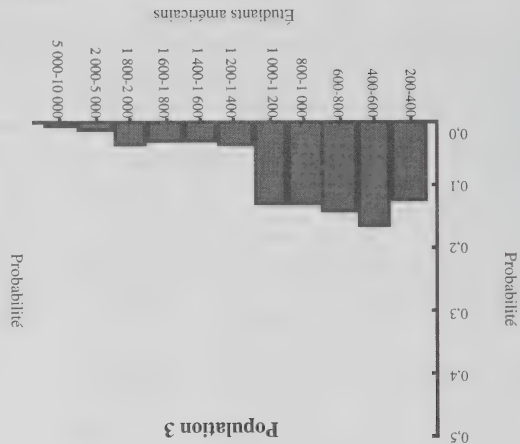
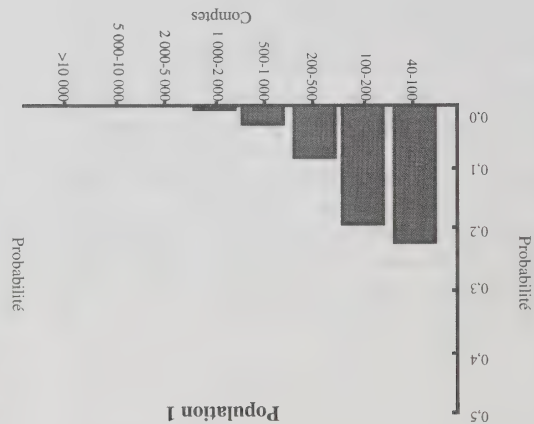
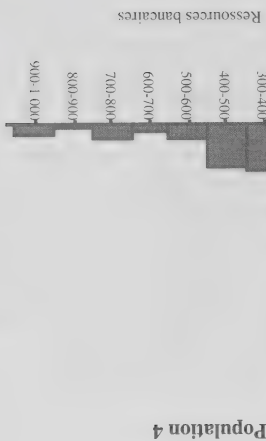
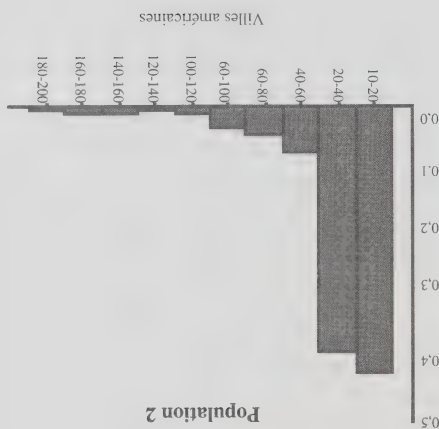


Figure 1. Populations



Hidiroglou.

Nous appliquons le nouvel algorithme à ces populations d'asymétrie. Les quatre populations utilisées sont illustrées et résumé à la figure 1 et au tableau 1 par ordre décroissant de la part supérieure de la population et disparaît si l'on élimine cette part 0,05 % de la population et disparaît si l'on élimine cette part

essentiellement discrète. Enfin, dans le dernier cas, une population de déclarations de revenus n'est pas suffisamment asymétrique; l'asymétrie est due à la part supérieure de la population de déclarations de revenus n'est pas suffisamment asymétrique.

sont souvent approximativement les mêmes pour toutes les strates. Il conclut cependant que calculer les écarts-types des strates et les égaliser est un exercice trop compliqué pour être réalisable en pratique. Dans la suite, nous montrons que, pour les distributions asymétriques, on peut égaliser approximativement les coefficients de variation entre les strates en recourant à la progression géométrique. Nous dérivons ce nouvel algorithme à la section 2. À la section 3, nous comparons l'efficacité de la nouvelle approximation à celle de la fonction cumulative de la racine carrée des fréquences et à celle de Lavallée et Hidiroglou. Enfin, nous résumons nos observations à la section 4.

2. UNE MÉTHODE DE RECHANGÉ POUR LA CONSTRUCTION DE STRATES

Stratifier une population selon la taille consiste à la subdiviser en intervalles dont les extrêmes sont $k_0 < k_1 < \dots < k_L$. Idéalement, la subdivision devrait être fondée sur la variable d'enquête X . Malheureusement, ce genre de construction est impossible, puisque cette variable X est inconnue; si nous la connaissons, nous n'aurions pas besoin de l'estimer. En pratique, nous utilisons donc une variable auxiliaire connue X , qui est corrélée à la variable d'enquête. Afin de déterminer les coupures (k_0, k_1, \dots, k_L) pour tous k_0 et k_L , nous cherchons à égaliser les $CV_h^* = S_{xh} / X_h$ pour $h = 1, 2, \dots, L$:

$$\frac{S_{x1}}{S_{x2}} = \frac{X_1}{X_2} = \dots = \frac{X_L}{S_{xL}}. \quad (3)$$

Maintenant, S_{xh} est l'écart-type et X_h la moyenne de X dans la strate h . Si nous émettons l'hypothèse que, dans chaque strate, la variable est approximativement uniforme-

$$X_h \approx \frac{k_h + k_{h-1}}{2},$$

$$S_{xh} \approx \frac{\sqrt{12}}{1} (k_h - k_{h-1}). \quad (5)$$

En tant qu'approximation des coefficients de variation, ceci nous donne

$$CV_h^* \approx \frac{(k_h - k_{h-1}) / \sqrt{12}}{(k_h + k_{h-1}) / 2}$$

et, si les CV_h^* sont égaux, nous devons avoir

$$\frac{k_{h+1} - k_h}{k_h + k_{h-1}} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}}. \quad (7)$$

Cette relation de récurrence nouvelle et exotique se réduit cependant à une forme familière :

$$k_h^2 = k_{h+1} k_{h-1}; \quad (8)$$

les bornes de stratification sont les termes d'une progression géométrique.

$$k_h = ar^h \quad (h = 0, 1, \dots, L). \quad (9)$$

Par conséquent, $a = k_0$, la valeur minimale de la variable, et $ar^L = k_L$, la valeur maximale de la variable. Il s'ensuit que le rapport constant peut se calculer sous la forme $r = (k_L / k_0)^{1/L}$. À titre d'exemple numérique, prenons $L = 4$; $k_0 = 5$; $k_4 = 50\,000$:

$$L = 4; \quad k_0 = 5; \quad k_4 = 50\,000; \quad (10)$$

donc, $k_h = 5 \cdot 10^{h/4}$ ($h = 0, 1, 2, 3, 4$) et les strates forment les intervalles

$$5 - 50; 50 - 500; 500 - 5\,000; 5\,000 - 50\,000. \quad (11)$$

Il s'agit manifestement d'une méthode fort simple d'obtention des bornes de stratification.

La relation (8) dépend de l'hypothèse selon laquelle les loïs de distribution à l'intérieur des strates sont uniformes, ce que l'on peut justifier grâce à l'argument heuristique qui suit. Lorsque la loi mère est positivement asymétrique, alors les valeurs faibles de la variable ont une forte incidence, qui diminue à mesure que les valeurs augmentent, si bien qu'il est approprié de prendre des intervalles plus petits au début de la distribution et des intervalles plus grands à la fin. C'est exactement ce qu'il se passe dans le cas d'une série géométrique dont le rapport constant est supérieur à 1. Dans la tranche des valeurs plus faibles de la variable, les strates sont étroites, si bien que l'hypothèse qu'elles suivent une loi rectangulaire n'est pas absurde. À mesure que la valeur de la variable augmente, la largeur de la strate s'accroît géométriquement. Ceci coïncide avec le taux de diminution de la variation de l'incidence de la variable positivement asymétrique, de sorte qu'ici également, l'hypothèse d'uniformité est raisonnable.

Cet algorithme ne fonctionne naturellement pas pour les loïs normales. En outre, comme les points de coupure augmentent géométriquement, il ne donne pas de bons résultats pour les variables pour lesquelles le point de départ est très faible, parce que le nombre de petites strates est trop grand; la règle ne fonctionne plus du tout lorsque la borne inférieure est nulle. En principe, les meilleurs résultats devraient être obtenus lorsque la distribution est très positivement asymétrique et que la partie supérieure ne contient qu'un faible pourcentage de la fréquence totale.

3. PERFORMANCES DE L'ALGORITHME

3.1 Certaines populations réelles positivement asymétriques

Pour tester notre algorithme, nous l'appliquons à quatre populations particulières, dont la distribution est asymétrique avec une queue positive.

La première (population 1) est une population de comptes débiteurs d'une entreprise irlandaise décrite en détail dans Horgan (2003). Nous utilisons aussi trois

Un nouvel algorithme pour la construction de bornes de stratification
dans les populations asymétriques

PATRICIA GUNNING et JANE M. HORGAN¹

RESUME

Nous devons un algorithme simple et pratique permettant d'établir des bornes de stratification telles que les coefficients de variation soient égaux dans chaque strate pour des populations positivement asymétriques. Nous montrons que, comparativement à la méthode de la fonction cumulative de la racine carrée des fréquences (Dalenius et Hodges 1957) et à la méthode d'approximation de L'avalée et d'Hindijoglu (1988), le nouvel algorithme est avantageux pour l'estimation des bornes optimales de stratification.

MOTS CLÉS : Efficacité; progression géométrique; répartition de Neyman; stratification.

1. INTRODUCTION

Un plan d'échantillonnage aléatoire simple stratifié est un plan d'échantillonnage en vertu duquel une population est subdivisée en strates mutuellement exclusives et des échantillons aléatoires simples sont tirés indépendamment à partir de chaque strate. L'objectif fondamental de la stratification est de construire des strates permettant une estimation plus efficace. Dans la suite de l'exposé, X représente la variable de stratification ou variable auxiliaire connue, tandis que Y représente la variable d'intérêt inconnue. Supposons qu'il existe L strates, contenant N_h éléments parmi lesquels un échantillon de taille n_h est sélectionné indépendamment de chaque strate ($1 \leq h \leq L$). Nous écrivons $N = \sum_{h=1}^L N_h$ et $n = \sum_{h=1}^L n_h$. Pour estimer la moyenne stratifiée,

$$(I) \quad {}^y\underline{\mathcal{L}} \frac{N}{yN} \sum_{\tau}^{l=y} = {}^{ys}\underline{\mathcal{L}}$$

où \bar{y}_h est la moyenne des éléments de l'échantillon dans la h^e strate, nous devons choisir les coupures de façon à minimiser sa variance

minimiser sa variance

$$(7) \quad S_z^y \left(\frac{y}{y} N - 1 \right) \left(\frac{N}{y} \right) \sum_l^1 = \binom{N}{y} \Lambda$$

no

$${}^uN/\left({}^uA-{}^{u'}A\right)\sum_{q_N}^{l=1}1={}_uS$$

est l'écart-type de Y limité à la strate h , et

$${}^{u}X \sum_{i=1}^I \frac{{}^iN}{I} = {}^uX$$

est la moyenne.

Dans le présent article, nous proposons un algorithme nettement plus simple à appliquer que n'importe lequel des algorithmes disponibles à l'heure actuelle. Il est fondé sur une observation de Cochran (1961), selon laquelle, dans le cas de bornes quasi optimales, les coefficients de variation

Dalenius (1950) a dérivé des équations en vue de déterminer les bornes de stratification lors de la stratification de variables selon la taille de façon à ce que l'expression (2) soit minimisée, mais ces équations se sont avérées difficiles à résoudre, à cause des dépendances entre les composantes. Depuis, de nombreux efforts ont été faits en vue d'obtenir des approximations efficaces de cette solution optimale. La première de ces approximations, proposée par Dalenius et Hodges (1957, 1959), consiste à construire les strates en prenant des intervalles égaux sur la fonction cumulative de la racine carrée des fréquences; cette méthode est encore utilisée aujourd'hui. La règle d'Eckman (1959), qui consiste à égaliser itérativement les produits du poids et de l'étendue des strates, demande des calculs ardu et est utilisée moins fréquemment que la méthode de Dalenius et Hodges (Nicholls 2001). Lavallée et Hidroglou (1988) ont élaboré une procédure itérative de stratification des populations asymétriques produisant une strate à tirage complet et un certain nombre de strates à tirage partiel, telles que la taille d'échantillon soit minimisée pour un niveau donné de fiabilité. D'autres contributions récentes incluent Hedlin (2000) qui a réexaminé la règle d'Eckman, Dorfman et Valliant (2000) qui ont comparé l'échantillonnage stratifié rond sur un modèle à l'échantillonnage équilibré, et Rives (2002) qui a construit une généralisation de l'algorithme de Lavallée et Hidroglou en fournissant des modèles tenant compte de l'écart entre la variable de stratification et la variable d'enquête.

BIBLIOGRAPHIE

BIEMER, P., et BUSHERY, J. (2001). Validité de l'analyse markovienne de structure latente pour l'estimation de l'erreur de classification des données sur la population active. *Techniques d'enquête*, 26, 2, 157-171.

BIEMER, P.P., et FORSMAN, G. (1992). On the quality of reinterview data with applications to the current population survey. *Journal of the American Statistical Association*, 87, 420, 915-923.

erreur-type n'a été fournie pour les estimations du tableau 2 et l'hypothèse d'un biais global plus faible dans le questionnaire révisé n'a pas été testée formellement. Malgré ces limites, les résultats semblent indiquer que la nouvelle série de données sur le chômage pourrait présenter un biais net considérablement plus faible que l'ancienne.

Tableau 2

Comparaison des biais dus aux questionnaires original et révisé pour le taux de chômage fondée sur les estimations d'après la CPS et l'Enquête parallèle de 1993

p	π	b_M	b_S	b_T
CPS de 1993	6,38	7,09	-0,71	-0,94
Enquête parallèle	6,98	8,03	-1,05	0 ¹
				-1,05

Nota : Il est supposé que le biais dû à l'erreur de spécification est nul pour les questions révisées.

ERREUR DE SPÉCIFICATION ET ERREUR DE

MESURE

Enfin, j'aborderai une question importante soulevée par M. Tucker en ce qui concerne l'erreur de spécification, l'erreur de mesure et leurs effets nets. Comme il l'explique, le questionnaire original souffrait d'un biais dû à une erreur de spécification causée par la mesure d'un concept incorrect. La révision des questions sur la population active introduite en 1994 avait pour but d'éliminer le biais dû à l'erreur de spécification grâce à une meilleure définition des concepts d'emploi et de chômage, et à la modification des questions d'enquête pour refléter ces perfectionnements. Si elles ont réduit l'erreur de spécification, ces modifications ont rendu les questions d'enquête plus complexes, ce qui pourrait avoir accru le biais dû à l'erreur de mesure dans les estimations concernant la population active. Selon M. Tucker, bien qu'il en soit peut-être ainsi, le biais de mesure dans la nouvelle série de données sur l'emploi pourrait être inférieur à la combinaison du biais de spécification et du biais de mesure présents dans l'ancienne série. Pour vérifier sa thèse, nous avons estimé séparément le biais dû à l'erreur de spécification (B_S) et le biais dû à l'erreur de mesure (B_M) au moyen des estimations MLCA présentées dans l'article, de la façon suivante.

Soit p l'estimation du chômage (UEM) selon la CPS et soit P l'espérance de p par rapport à l'échantillonnage et aux distributions des erreurs de mesure. Soit π la valeur réelle de la caractéristique sous la définition du chômage impliquée par le questionnaire dont il est question (c'est-à-dire sans tenir compte de l'erreur de spécification éventuelle). Par conséquent, $\pi = P - B_M$, c'est-à-dire la valeur de P en l'absence de biais dû à l'erreur de mesure.

Tel que mentionné plus haut, le biais dû à l'erreur de spécification est le biais dans P causé par l'utilisation d'un concept ou d'une définition du chômage incorrect découlant des questions et (ou) du processus de classification de la population active. Dans le cas du questionnaire révisé, nous supposons que l'erreur de spécification dans P est nulle, puisque ce questionnaire sera considéré comme étalon pour estimer le biais dû à l'erreur de spécification dans le questionnaire original.

Soit π_{original} et $\pi_{\text{révisé}}$ le paramètre π pour les questionnaires original et révisé, respectivement. Alors, le biais dû à l'erreur de spécification dans les estimations du taux de chômage antérieures à 1994 est

$$B_S = \pi_{\text{original}} - \pi_{\text{révisé}} \quad (2)$$

provenant de la MLCA. Soit $\mathbf{p}' = (p_1, p_2, p_3)$, où p_1, p_2, p_3 représentent les estimations des proportions de personnes dans les catégories des personnes occupées (EMP), en chômage (UEM) et inactives (NLF), respectivement. Soit w_j la probabilité qu'une observation appartenant à la i^{e} catégorie soit attribuée à la j^{e} catégorie et soit π_j la proportion réelle de la population dans la j^{e} catégorie. Alors

$$E(\mathbf{p}) = \mathbf{\Sigma} \pi \quad (3)$$

où $\pi = (\pi_1, \pi_2, \pi_3)'$ et $\mathbf{\Sigma} = [w_{ij}]$ est la matrice de dimensions 3×3 des éléments w_{ij} . Il s'ensuit qu'un estimateur de π est

$$\hat{\pi} = (\hat{\Sigma})^{-1} \mathbf{p} \quad (4)$$

où $\hat{\Sigma}$ est une estimation MLCA de $\mathbf{\Sigma}$. Pour chaque questionnaire, nous avons estimé $\hat{\Sigma}$ par la moyenne des 10 estimations MLCA (de janvier à mars jusqu'à octobre à décembre) au moyen des données de la CPS de 1993 pour le questionnaire original et de celles de l'Enquête parallèle de 1993 pour le questionnaire révisé.

Le tableau 2 montre les résultats de cette analyse. Pour la catégorie UEM, $p = 6,38$ pour le questionnaire original et 6,98 pour le questionnaire révisé. Si l'on corrige les taux de chômage pour le biais de mesure au moyen de (4), le taux de chômage augmente pour atteindre 7,09 % pour le questionnaire original et 8,03 % pour le questionnaire révisé. Donc, l'estimation du biais de mesure pour le questionnaire original est 6,38 - 7,09 = - 0,71 et pour le questionnaire révisé, 6,98 - 8,03 = - 1,05. Il convient de souligner que les biais de mesure sont négatifs tant pour le questionnaire original que pour le questionnaire révisé, ce qui indique aussi que le taux de chômage est sous-estimé d'après les deux versions du questionnaire.

Pour le questionnaire révisé, nous supposons que le biais de spécification est nul. Pour le questionnaire original, ce biais est estimé par différence, soit 7,09 - 8,03 = - 0,94 %. L'estimation du biais net, $B_T = B_M + B_S$, est - 0,71 + (- 0,94) = - 1,65 % pour l'ancienne série comparativement à - 1,05 + 0 = - 1,05 % pour la nouvelle. Donc, bien qu'elle présente un biais dû à l'erreur de mesure plus important, la nouvelle série est caractérisée par un biais net estimatif plus faible, si l'on suppose que $B_S = 0$.

Plusieurs limites de ces résultats doivent être mentionnées. Premièrement, comme nous l'indiquons dans l'article d'après l'Enquête parallèle pour le questionnaire révisé principal, les estimations pour le questionnaire révisé représentatives de la série révisée de la CPS. Deuxièmement, l'analyse se fonde sur l'hypothèse que le questionnaire révisé est l'étalon d'or pour l'estimation du biais dû à l'erreur de spécification dans le questionnaire original. Cette hypothèse pourrait aussi être contestée. Enfin, aucune

markovienne. Il est donc peu probable que des transitions non markoviennes expliquent l'observation d'une erreur de classification plus grande pour le questionnaire révisé. Néanmoins, nous devons poursuivre les travaux en vue de mieux comprendre les effets des transitions non

markoviennes sur nos résultats.

En ce qui concerne (b), il est fort possible que les estimations MLCa soient biaisées si la composition de la population de chômeurs diffère considérablement sous les questionnaires original et révisé, et que les différences ne

sont pas expliquées par les variables de groupement utilisées dans le modèle. Par conséquent, (c) peut être considéré comme un cas particulier de (b). En ce qui concerne (c), les probabilités de transition pour un sous-groupe donné de la population ne sont pas corrélées à la situation d'activité le mois précédent; au contraire, elles sont corrélées à d'autres variables *inobservées*. Dans l'exemple du buveur de café de Jerroen Vermunt, la variable *inobservée* est la disponibilité d'une marque particulière de café au supermarché. À ce stade de la recherche, nous n'avons pas réalisé d'étude en simulation afin de quantifier les effets de l'hétérogénéité *inobservée* sur les estimations, mais nous envisagerons cette possibilité dans l'avenir.

Cependant, cette question et celle de la vraisemblance générale des estimations MLCa peuvent être étudiées dans une certaine mesure en comparant les estimations MLCa à des estimations indépendantes, obtenues selon une approche sur laquelle les points (a) à (c) n'ont aucune incidence. Si les résultats de cette autre analyse concordent avec ceux de la MLCa, ces derniers seront plus crédibles. Par exemple, on peut estimer la fiabilité test-rétest des classifications de la situation d'activité selon la CPS avant et après le remaniement en utilisant les données de réinterview de la CPS (voir, par exemple, Biemer et Forsman (1992) pour une description du programme de réinterview de la CPS et de ces données). La validité des estimations de la fiabilité test-rétest ne dépend pas de l'hypothèse markovienne ni de l'hypothèse d'homogénéité des groupes; par contre, l'hypothèse d'IBC demeure pertinente pour l'estimation de la fiabilité.

Le tableau 1 contient les estimations du coefficient kappa de Cohen donnant la mesure de fiabilité pour trois périodes, à savoir 1992 – 1993, 1995 – 1997 et 2002 – 2003. Comme le montre le tableau, la fiabilité de la classification des personnes en chômage selon la CPS a diminué après le remaniement, pour passer d'environ 68 % à 65 %. Les estimations les plus récentes du coefficient kappa indiquent que le niveau de fiabilité est tombé sous la barre des 60 %. Ces résultats concordent avec ceux de la MLCa selon lesquels l'erreur de classification dans la catégorie des personnes en chômage selon la CPS s'est aggravée après le remaniement. Il se peut que les estimations de la fiabilité du tableau 1 soient biaisées, puisqu'elles s'appuient aussi sur la

validité de l'hypothèse d'IBC. Cependant, tel que discuté plus haut, pour pouvoir expliquer les résultats du tableau par la violation de l'hypothèse d'IBC, il faudrait que cette dernière tienne pour les questions révisées, mais non pour les questions originales. Or, il est peu probable que cette condition soit satisfaite.

Tableau 1

Estimations du coefficient kappa de Cohen pour la CPS avant et après le remaniement

Année	n	k de Cohen
1992 – 1993 ¹	28 063	67,8
1995 – 1997 ²	22 429	64,6
2002 – 2003 ³	19 205	58,8

¹ Biemer et Bushery (2000).

² Bushery et McGovern (1999).

³ Communication personnelle avec Bac Tran du U.S. Census Bureau.

Compte tenu des données présentées ici et dans l'article principal, il semble raisonnable de penser que l'erreur de classification dans la catégorie de personnes en chômage de la CPS pourrait avoir augmenté après le remaniement. L'étape suivante consistera à mener une étude approfondie en vue d'évaluer ces résultats et d'explorer les causes éventuelles de l'erreur. Au lieu de se concentrer sur la validité des modèles MLCa ou des modèles de réinterview test-rétest, les futurs travaux de recherche devraient porter sur les questions révisées de la CPS, particulièrement celles utilisées pour la classification dans la catégorie des personnes mises à pied (LAYOFF). J'ai déjà mentionné la possibilité d'utiliser des interviews cognitives pour détecter les problèmes posés par le processus de réponse qui sont associés aux questions révisées. Par exemple, une question qui, selon la MLCa, pourrait être imparfaite est : « Vous a-t-on donné des raisons de croire que vous seriez rappelé(e) au travail d'ici six mois? » Voici certains problèmes concernant cette question susceptibles d'être examinés au laboratoire d'analyse cognitive :

– Dans quelle mesure les sujets en chômage comprennent-ils bien la signification des termes « des raisons de croire » et « rappelé(e) »?

– Les sujets dont l'emploi s'est terminé récemment ont-ils des difficultés à se souvenir de ce que leur employeur leur a dit au sujet d'un rappel au travail au moment où leur emploi a pris fin?

– Un employeur pourrait dire : « Si les affaires vont mieux, nous pourrions vous rappeler. » Les répondants répondent-ils à la question correctement dans cette situation?

– Les répondants qui ont déclaré au départ qu'ils seraient rappelés au travail plus tard modifient-ils leur réponse à cette question à mesure que les mois s'écoulent et qu'ils n'ont pas été rappelés.

variant en fonction du temps, ainsi que d'autres améliorations mentionnées par M. Vermunt.

COMMENTAIRES DES CRITIQUES DU BLS

J'aborderai ensemble les commentaires de M. Miller et Mme Polivka, ainsi que ceux de M. Tucker, puisque ces examinateurs appartiennent au même organisme (BLS) et qu'ils émettent des réserves comparables au sujet de l'analyse. Les cinq points qui suivent résument leurs principales préoccupations :

1. Les modifications introduites dans le nouveau questionnaire permettent de repérer un plus grand nombre de transitions qu'au moyen de l'ancien questionnaire. L'analyse MLCA interprète cette augmentation du nombre de transitions repérées comme étant des erreurs, quand, en fait, il n'en n'est rien.

2. L'opinion des répondants quant au fait que leur employeur leur a vraiment indiqué qu'ils pourraient être rappelés au travail peut changer de mois en mois. Ces changements ne devraient pas être considérés comme une erreur de réponse.

3. L'hypothèse markovienne ne tient pas dans les études de la population active et la violation est encore plus importante après le remaniement qu'avant celui-ci. Cette violation différentielle des hypothèses du modèle pourrait influencer fondamentalement les résultats de la MLCA.

4. Les écarts entre les estimations de l'erreur de classification pour la catégorie des personnes mises à pied (LAYOFF) avant et après le remaniement sont dus à la composition des groupes classés dans cette catégorie. Cette composition a changé après le remaniement d'une façon qui était souhaitée et voulue par les auteurs du remaniement du questionnaire.

5. La discordance croissante entre les réponses aux questions sur la recherche d'emploi (LOOKING) observée pour les questions révisées pourrait être due au fait qu'un plus grand nombre de travailleurs marginaux sont dépistés à l'aide des questions révisées. À un certain moment, ces individus sont réellement à la recherche de travail et à d'autres, non. La MLCA interprète incorrectement ces variations manifestement aléatoires comme étant une erreur de réponse quand ce n'est pas le cas.

Le point 1 décrit un problème qui ne devrait poser aucune difficulté pour la MLCA. Le modèle MLCA repose sur l'hypothèse que chaque individu occupe un état

d'activité réel qui peut changer de mois en mois. Il n'est pas supposé que les probabilités de transition sont les mêmes pour les deux questionnaires. Les probabilités initiales réelles des états d'activité, ainsi que les probabilités de transition d'un mois à l'autre sont estimées indépendamment pour chaque questionnaire. En fait, bien que ce point ne soit pas discuté dans l'article principal, les estimations des probabilités réelles de sortie des états LOOKING et LAYOFF fondées sur les modèles sont en fait plus grandes pour le questionnaire révisé que pour le questionnaire original. Donc, un plus grand nombre de flux d'une catégorie d'activité à une autre pour le questionnaire révisé n'introduit pas nécessairement un biais dans les estimations de l'erreur de classification pour cette catégorie.

Le point 2 laisse entendre que le fait qu'un individu soit vraiment mis à pied dépend du fait que cet individu pense ou non qu'on lui a indiqué qu'il pourrait être rappelé au travail. Cependant, ce n'est pas ainsi que le questionnaire révisé définit le concept. L'état réel de mise à pied d'un individu dépend du fait que l'employeur lui a ou non vraiment donné des raisons de croire qu'il serait rappelé au travail. Bien que l'opinion qu'a le répondant de ce que l'employeur a indiqué puisse évoluer de mois en mois, l'état réel de mise à pied ne varie pas selon l'opinion du répondant. Les flux d'entrée dans la catégorie LAYOFF et les flux de sortie de cette catégorie dus à l'opinion du répondant devraient être interprétés comme une erreur par le modèle.

Les points 3, 4 et 5 pourraient être formulés pour toute analyse fondée sur la MLCA. Ils ont essentiellement trait au biais éventuel dans les estimations MLCA lorsque les transitions de mois en mois n'ont pas lieu conformément au modèle MLCA et, conséquemment, que les variations réelles sont interprétées incorrectement comme des erreurs de classification. Comme le font remarquer les examinateurs, cette situation peut se produire au moins de trois façons :

- a) L'hypothèse markovienne n'est pas vérifiée (point 3);
- b) la population présente une hétérogénéité inobservée et inexpliquée (point 4);
- c) les comportements liés à l'emploi pendant deux mois consécutifs ne sont pas corrélés chez certaines personnes; donc, pour ces personnes, la situation du mois précédent ne prédit pas la situation du mois courant (point 5, ainsi qu'une remarque formulée par M. Vermunt).

Les implications de (a) ont été considérées dans une analyse en simulation décrite dans Biemer et Bushery (2001). Leurs résultats donnent à penser que, pour les données de la CPS, les estimations de l'erreur de classification sont assez robustes aux violations de l'hypothèse

violations des hypothèses du modèle MLCA, ils ne fournissent aucune preuve directe de la validité des estimations produites d'après ce modèle. Biemer et Bushery (2001) montrent comment on peut établir la validité (empirique) des estimations de classes latentes à l'aide de données externes, en recourant à d'autres approches pour estimer l'erreur de classification. Une analyse semblable fondée sur les données de réinterview (test-retest) sera présentée dans la suite.

Afin de relever les domaines dans lesquels le questionnaire de la CPS pourrait être amélioré, il n'est pas essentiel d'établir catégoriquement que les hypothèses du modèle MLCA sont vérifiées, puisque la validité du modèle est une question d'importance secondaire. Au contraire, la question principale à laquelle doivent répondre les travaux d'évaluation du questionnaire est celle de savoir si la méthode d'analyse utilisée permet de repérer les questions qui produisent d'importantes erreurs de mesure et ont besoin d'être révisées. Autrement dit, la capacité qu'a le modèle de dépisier d'importants défauts dans le questionnaire établi sa validité. Déterminer s'il existe vraiment une erreur dans la classification des personnes en chômage (UEM), comme le suggère l'approche MLCA, nécessite une évaluation selon d'autres méthodes, telle que l'étude cognitive en laboratoire. On pourrait recourir à des interviews cognitives pour examiner des problèmes de codage, de compréhension, de remémoration et (ou) de désirabilité sociale qui génèrent des erreurs dans les réponses aux questions sur le chômage. Si ces investigations révèlent des problèmes importants dans les questions, alors l'utilité de la MLCA pour repérer les questions imparfaites sera corroborée, même si la validité des hypothèses de modélisation MLCA n'est jamais établie.

Les autres propositions de M. Vermaun quant à la façon d'améliorer le cadre de modélisation sont raisonnables et j'espère les étudier de façon plus approfondie dans l'avenir. Cependant, le logiciel utilisé à l'heure actuelle pour ajuster les modèles MLCA présente certaines limites et l'estimation de modèles complexes, tels que ceux qu'il propose, pourrait ne pas être faisable. Il note aussi que des problèmes peuvent se poser lors de l'ajustement de grands modèles au moyen de l'algorithme EM. Par exemple, au départ, nous avons essayé d'utiliser la variable de réponse par procuration/sans procuration comme covariable variant en fonction du temps dans les modèles MLCA, mais nous sommes heurtés à des obstacles dans le processus d'estimation, comme des erreurs de « division par 0 » et la convergence persistante vers les maxima locaux. En dernière analyse, nous avons dû abandonner l'approche au profit de la variable de groupement unique « par procuration/sans procuration », invariante dans le temps. À mesure que de nouveaux logiciels plus généraux seront disponibles, il sera possible d'utiliser des options de modèle MLCA avec covariables

Suivant ses recommandations, nous avons réalisé une petite étude en simulation pour mieux comprendre les conséquences de $p > 0$ pour la MLCA en utilisant les données de la CPS. Nous avons généré une série de populations artificielles en utilisant des paramètres concordant avec ceux de la CPS (voir, par exemple, le tableau 1 de l'article principal), à part le fait que la valeur de p a été augmentée progressivement pour passer de 0 à la valeur empirique maximale, c'est-à-dire la valeur la plus grande de p applicable sans violer les autres hypothèses du modèle. Il est, en effet, nécessaire de retenir les autres hypothèses du

Nous avons établi empiriquement que la plus grande valeur applicable de p était 0,7. À cette valeur de p , l'estimation par la MLCA de la probabilité que la classification dans la catégorie des personnes en chômage (UEM) soit correcte passe de 79 % à 85 % et le taux d'erreur de classification baisse, pour passer de 21 % à 15 %. Pour les faibles écarts par rapport à l'hypothèse d'IEC, disons $0 < p < 0,3$, les taux d'erreur varient de moins de trois points. Ces résultats montrent que, si l'hypothèse d'IEC ne tient pas à cause de corrélations positives entre les interviews, les taux d'erreur estimés par la MLCA seront légèrement sous-estimés. Cependant, des écarts faibles par rapport à l'hypothèse d'IEC ne devraient avoir que peu d'effets sur les probabilités d'erreur de classification pour ces données. Une analyse semblable a été réalisée pour les deux autres catégories d'activité (c'est-à-dire les personnes occupées (EMP) et les personnes inactives (NLF)), mais la variation des estimations de l'erreur de classification a été négligeable. Ce résultat était prévu, étant donné les taux d'erreur relativement faibles pour ces catégories.

Les résultats donnent à penser que de légers écarts par rapport à l'hypothèse d'IEC ne devraient avoir que peu d'effet, voir aucun, sur les conclusions de l'analyse. Des écarts très importants pourraient influencer les conclusions dans le cas peu probable où les erreurs seraient fortement corrélées pour le questionnaire original et essentiellement non corrigées pour le questionnaire révisé. Sous ce scénario, le questionnaire original semble produire une erreur de classification plus faible que le questionnaire révisé pour la catégorie des personnes en chômage (UEM). Cependant, il n'existe aucune raison pratique de s'attendre à ce que cette condition soit vérifiée, puisque les deux questionnaires contiennent des questions que les répondants pourraient mal comprendre systématiquement lors des interviews réussies.

Bien que ces résultats de simulation, ainsi que ceux de Biemer et Bushery (2001) pour l'étude des conséquences de la violation de l'hypothèse markovienne, soient relativement utiles pour l'étude de la sensibilité des estimations aux

Réponse de l'auteur

PAUL P. BIEMER¹

COMMENTAIRES DE JEROEN VERMUNT

J'examinerai d'abord les commentaires de M. Vermunt, puis ceux des trois autres critiques. Comme M. Vermunt, je pense que l'hypothèse d'indépendance des erreurs de classification (IEC) pourrait ne pas tenir dans le cas des données analysées. Comme il le souligne, si les répondants comprennent mal les questions sur la population active de la même façon d'un mois à l'autre, ils risquent de commettre les mêmes erreurs chaque mois, si bien que les erreurs seront corrélées. Par exemple, une personne qui appartenait réellement à la catégorie des personnes en chômage (UEM) aux périodes 1 et 2 est plus susceptible d'être classée erronément à la période 2 si elle a également été mal classée à la période 1. De façon probabiliste, ceci peut s'énoncer comme suit

$$p = \frac{P(B \neq 2 | A \neq 2 \text{ et } X = Y = 2)}{P(B \neq 2 | A = 2 \text{ et } X = Y = 2)} - 1 > 0. \quad (1)$$

La probabilité figurant au numérateur de la quantité p est la probabilité que la classification à la période 2 (B) soit une erreur sachant que la classification à la période 1 (A) est également une erreur et que la classification correcte aux deux points dans le temps est UEM (personnes en chômage). La probabilité figurant au dénominateur est semblable, à part la condition qu'aucune erreur n'est faite à la période 1 (c'est-à-dire $A = 2$). Sous l'hypothèse d'IEC, $p = 0$. Par conséquent, si $p > 0$ (ce qui est la direction probable de l'erreur corrélée), l'hypothèse d'IEC est violée. M. Vermunt propose de réaliser une étude en simulation pour déterminer la sensibilité des estimations de l'erreur de classification à la violation de cette hypothèse. Naturellement, il est impossible de déterminer par simulation dans quelle mesure l'hypothèse d'IEC ne tient pas pour les données de la CPS. Il n'en reste pas moins utile d'évaluer la possibilité que la corrélation des erreurs introduise un biais dans les estimations de l'erreur de classification par la MLCA.

Je remercie sincèrement les quatre critiques de leurs commentaires réfléchis, complets et constructifs qui nous ont permis d'accroître considérablement notre compréhension des questions complexes que soulèvent l'analyse markovienne de classes latentes (MLCA) et les statistiques sur la population active produites d'après la Current Population Survey (CPS). Ils soulèvent un certain nombre de questions importantes auxquelles je vais essayer de répondre de mon mieux. Certaines nécessitent une étude plus approfondie et méritent un exposé plus détaillé qu'il n'est possible ici. Il faudra attendre les résultats de futurs travaux de recherche pour y répondre complètement.

Considérés collectivement, les commentaires indiquent que les critiques semblent être d'accord que l'analyse markovienne de classes latentes offre de grandes possibilités en tant qu'outil d'évaluation et d'exploration des sources d'erreur de mesure dans la CPS. Cependant, ils paraissent douter qu'elle ait permis de déceler des problèmes réels dans le questionnaire de la CPS. M. Vermunt, qui est également l'auteur du logiciel utilisé pour l'analyse (c'est-à-dire EEM), a formulé plusieurs suggestions intéressantes en vue d'améliorer les modèles et d'étudier la validité des hypothèses qui les sous-tendent. Les trois autres examinateurs (M. Miller, Mme Polivka et M. Tucker) connaissent bien la CPS, puisqu'ils travaillent pour l'organisme fédéral qui parraine l'enquête et qu'ils ont joué un rôle important dans le remaniement de 1994. Leurs commentaires démontrent les diverses façons dont les hypothèses du modèle MLCA pourraient être violées dans le cas des données de cette enquête. En outre, ils contiennent des renseignements précieux sur la CPS (avant et après le remaniement) et la création de variables de population active fondées sur cette enquête. Les commentaires et suggestions de ces quatre critiques devraient être considérés minutieusement par les économistes et les statisticiens spécialisés dans la population active qui poursuivent des travaux de recherche dans le domaine de l'erreur de mesure de l'emploi, particulièrement ceux qui utilisent la MLCA.

1. INTRODUCTION

BIBLIOGRAPHIE

- POLIVKA, A.E., et MILLER, S. (1998). The CPS after the redesign: Refocusing the economic lens. Dans *Labor Statistics Measurement Issues*. (Eds. J. Haltiwanger, M.E. Manser, et R. Topel). National Bureau of Economic Research Studies in Income and Wealth. Chicago: University of Chicago Sous presse, 60, 249-286.
- POLIVKA, A.E., et ROTHGER, J. (1993). Overhauling the Current Population Survey: Redesigning the questionnaire. *Monthly Labor Review*, 116, 10-28.
- MILLER, S.M., et POLIVKA, A.E. (2004). Commentaire de l'article Une analyse de l'erreur de classification pour les questions sur l'emploi révisées de la Current Population Survey. *Techniques d'enquête*, 30, 161-166.
- ESPOSITO, J.T., CAMPANELLI, P.C., ROTHGER, J.M., et POLIVKA, A.E. (1991). Determining which questions are best: Methodologies for evaluating survey questions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 46-55.
- BREGGER, J.E., et DIPPO, C.S. (1993). Overhauling the Current Population Survey: Why is it necessary to change? *Monthly Labor Review*, 116, 3-9.

fait le plus intéressant en ce qui me concerne, ces questions visent à recueillir des renseignements sur des changements assez nuancés. Par exemple, un répondant pourrait changer d'avis quant à la possibilité d'être rappelé au cours des six mois à venir en ne s'appuyant que sur peu de renseignements concrets. Compte tenu des incertitudes qui caractérisent le marché du travail aujourd'hui, il serait difficile d'affirmer que ce répondant a répondu incorrectement.

Je voudrais maintenant passer aux réserves de Biemer quant à la première question de la nouvelle série sur l'activité, à savoir si un travail a été fait la semaine précédente « contre rémunération ou en vue d'un bénéfice ». Selon lui, cette question pourrait contribuer à l'erreur qu'il observe dans les séries sur les « personnes mises à pied » et les « personnes à la recherche d'emploi ». La modification de cette question (ainsi que l'ajout d'une question sur l'existence d'une entreprise ou d'une ferme familiale) vise à répondre à la crainte que l'énoncé des anciennes questions n'était pas assez général, si bien que les travailleurs marginaux, particulièrement ceux travaillant à la maison en vue d'un bénéfice n'étaient pas classés dans la catégorie des personnes occupées. Par exemple, l'Enquête parallèle montre que le pourcentage de travailleurs à temps partiel selon la nouvelle CPS est supérieur d'un facteur de 1,09 à celui calculé d'après l'ancienne CPS et que, tout à fait par hasard, le ratio de l'emploi à la population pour les femmes de 65 ans et plus a aussi augmenté dans presque la même proportion (Polivka et Miller 1998). Il en est de même si l'on compare 1993 et 1994. Il va sans dire que la plus grande précision de l'identification de ces travailleurs marginaux, qui sont plus susceptibles que les autres de donner des réponses non concordantes de mois en mois, pourrait être considérée à tort comme une erreur de mesure. Le fait est que la question plus restreinte « Que faisiez-vous la semaine dernière » pourrait pousser les répondants à déclarer plus uniformément, mais inexactement, qu'ils étaient en chômage.

Enfin, j'examinerais l'autre section de la série de questions sur l'activité que Biemer a jugées problématiques, c'est-à-dire celles sur la « recherche d'emploi ». Dans cette série, un changement important vise à rendre plus claires les différences entre la recherche « active » et « inactive » d'un emploi afin de réduire le taux d'erreur de classification dans ces catégories. Deux études réalisées durant les années 1990 ont révélé que les intervieweurs avaient du mal à définir la recherche active (par opposition à passive) d'emploi (Polivka et Rohgeb 1993). Dans le questionnaire remanié, ils disposent d'une liste explicite de méthodes de recherche active et passive d'emploi.

Le fait que les anciennes et les nouvelles questions soient posées à des sous-populations différentes complique la comparaison des résultats obtenus. Les personnes classées, en bout de ligne, dans la catégorie des personnes à la recherche d'emploi, donc considérées comme étant en chômage, ont abouti dans cette catégorie de façon assez

4. CONCLUSION

différente. La moitié des personnes considérées comme étant à la recherche d'emploi en 1993 ont été classées de la sorte en indiquant volontairement qu'elles recherchaient un emploi en réponse à la première question (« Qu'avez-vous fait le plus la semaine dernière? »); aucune des personnes classées comme étant à la recherche d'emploi en 1994 n'ont suivi ce chemin. Les personnes à la retraite et celles âgées de 50 ans et plus en 1994 n'ont jamais eu la chance de dire qu'elles recherchaient un emploi. En 1993, la question sur la recherche d'emploi n'a été posée à aucune des personnes qui ont dit avoir été mises à pied, si bien qu'elles n'ont pas eu la chance d'être classées dans la catégorie des personnes inactives durant un mois particulier. En outre, les intervieweurs ont reçu des informations de niveaux différents pour coder les méthodes actives et passives. L'une des différences que révèle l'analyse des deux groupes de produits d'après les données de 1993 et celles de 1994 est qu'une plus forte proportion des personnes à la recherche d'un emploi étaient de sexe féminin en 1994 qu'en 1993 (45,4 % c. 41,2 %). En revenant à la discussion concernant la première question sur la situation d'activité, le plus grand manque de convergence des réponses aux questions sur la recherche d'emploi pourrait tenir au fait que la série de questions révisées permet de repérer un plus grand nombre de travailleurs marginaux, c'est-à-dire des personnes qui sont parfois à la recherche d'emploi et à d'autres moments pas.

La tentative de Paul Biemer en vue d'étudier la structure de l'erreur de la série de données sur l'activité de la CPS est audacieuse; cependant, ses résultats ne tiennent pas compte de l'erreur de la série de données sur l'activité de la CPS afin de tenir compte de certains changements méthodologiques et de l'erreur de mesure dans des sous-populations plus restreintes. Il conviendrait aussi de se livrer à un examen plus approfondi de l'utilité de la MLCA dans le cas de classifications intrinsèquement non convergentes.

REMERCIEMENTS

Les opinions exprimées dans le présent article sont celles de l'auteur et ne représentent pas une politique du Bureau of Labor Statistics. L'auteur remercie Steve Miller, Anne Polivka et John Dixon de leur aide au sujet de cette discussion.

[Traduction] « ... au début des années 1980, l'introduction de deux nouvelles méthodologies d'enquête a permis de comprendre et de réduire l'erreur de mesure. Il s'agit de l'application des méthodes et de la théorie des sciences du comportement, appelées plus communément aspects cognitifs de la méthodologie d'enquête, et de l'interview assistée par ordinateur. C'est grâce à l'amalgamation de ces deux méthodologies qu'une nouvelle procédure de collecte, axée sur la réduction de l'erreur de mesure, est devenue possible. »

Les méthodes cognitives (y compris les groupes de discussion et l'interview en profondeur) ont permis d'élaborer des questions produisant une mesure exacte des comportements économiques plus complexes caractérisques du temps présent. De surcroît, ces techniques ont permis de découvrir des problèmes dans la série existante sur la population active (voir Polivka et Rothgeb 1993). La mesure exacte des comportements plus complexes a aussi nécessité un questionnaire plus compliqué. Si compliqué que les intervieweurs, livrés à eux-mêmes, auraient eu de la difficulté à le maîtriser. C'est à cet égard que l'interview assistée par ordinateur a joué un rôle important. Doté d'un questionnaire informatisé, les intervieweurs ont pu suivre facilement les enchaînements de questions complexes en vue d'obtenir les réponses aux questions permettant de mesurer le grand éventail de comportements économiques étudiés.

3. EXAMEN DES ERREURS NON ATTRIBUABLES À L'ÉCHANTILLONNAGE DANS LES ANCIENNE ET NOUVELLE SÉRIES DE DONNÉES SUR L'ACTIVITÉ DE LA CPS

Je commencerai par exposer en détail les raisons pour lesquelles j'estime que la MLCA n'est pas un bon outil pour évaluer la nouvelle conception de la CPS comparativement à l'ancienne. La MLCA peut, certes, être une bonne méthode de dépistage de l'erreur de mesure dans une série constante de questions par la recherche des discordances entre les réponses fournies par un même répondant lors de plusieurs cycles d'enquête. Dans le cas de la CPS, la méthode pourrait convenir au dépistage indépendant des problèmes dans la nouvelle et dans l'ancienne série de données sur l'activité, à condition de se livrer à un examen minutieux d'un ensemble bien choisi de diagnostics. Toutefois, j'ajouterais ici une mise en garde concernant l'examen des discordances, même dans une même série de données sur l'activité, à condition de se livrer à un examen des discordances, même dans une même série de données sur l'activité, car la situation d'activité est intrinsèquement non homogène au cours du temps. Alors que les catégories des personnes occupées et des personnes inactives sont relativement stables, celle des personnes en

Manifesterment, ces changements ont modifié les caractéristiques du groupe classé dans la catégorie des personnes en chômage à la suite d'une mise à pied, ainsi que des personnes auxquelles sont posées les autres questions de la série sur l'activité, mais je pense qu'il existe des raisons plus subtiles pour lesquelles le manque de concordance des réponses pourrait s'être accentué, sans toutefois avoir contribué à l'erreur de classification dans la mesure où le prétend Biemer. En premier lieu, les enquêtes doivent augmenter la probabilité que survienne au moins une fausse discordance d'un mois à l'autre. Cette situation risque d'accroître l'erreur de mesure comparativement à l'ancienne série de questions, mais l'erreur de spécification, qui est considérée être le problème le plus important, serait malgré tout réduite. En outre, les fausses discordances dues à ces questions sont, en principe, réduites au minimum pour deux raisons. Les nouvelles questions sont beaucoup plus spécifiques que la question unique sur la mise à pied de l'ancienne série, et elles ont été bien testées (Esposito, Campanelli, Rothgeb et Polivka 1991). De surcroît, puisque des questions plus spécifiques sont posées, il est probable qu'un changement réel de l'état décrit par au moins l'une d'elles ait eu lieu durant le mois écoulé. Enfin,

Manifesterment, ces changements ont modifié les caractéristiques du groupe classé dans la catégorie des personnes en chômage à la suite d'une mise à pied, ainsi que des personnes auxquelles sont posées les autres questions de la série sur l'activité, mais je pense qu'il existe des raisons plus subtiles pour lesquelles le manque de concordance des réponses pourrait s'être accentué, sans toutefois avoir contribué à l'erreur de classification dans la mesure où le prétend Biemer. En premier lieu, les enquêtes doivent augmenter la probabilité que survienne au moins une fausse discordance d'un mois à l'autre. Cette situation risque d'accroître l'erreur de mesure comparativement à l'ancienne série de questions, mais l'erreur de spécification, qui est considérée être le problème le plus important, serait malgré tout réduite. En outre, les fausses discordances dues à ces questions sont, en principe, réduites au minimum pour deux raisons. Les nouvelles questions sont beaucoup plus spécifiques que la question unique sur la mise à pied de l'ancienne série, et elles ont été bien testées (Esposito, Campanelli, Rothgeb et Polivka 1991). De surcroît, puisque des questions plus spécifiques sont posées, il est probable qu'un changement réel de l'état décrit par au moins l'une d'elles ait eu lieu durant le mois écoulé. Enfin,

chômage ne l'est pas. Les personnes appartenant à cette dernière catégorie essaient d'en sortir. Si l'on neutralise les effets saisonniers en comparant la période de mars à mai de 1993 ou de 1994, il s'avère qu'en moyenne, près de 90 % des personnes occupées et des personnes inactives n'ont pas changé de catégorie d'un mois au suivant. Par contre, plus de la moitié des personnes en chômage l'ont fait. Donc, les chômeurs constituent un groupe qu'il est particulièrement difficile de traiter dans le cadre de la MLCA.

Commentaire
CLYDE TUCKER¹

1. INTRODUCTION

Je voudrais féliciter Paul Biemer d'offrir une approche novatrice de l'étude de l'erreur de mesure dans les enquêtes. Bien qu'il ait choisi de l'illustrer au moyen de la série de données sur l'activité provenant de la Current Population Survey (CPS), la méthode est applicable à de nombreuses enquêtes. Mes commentaires sont, en grande partie, de nature conceptuelle, mais je les compléterai par des exemples tirés des mêmes données que celles analysées par Biemer.

Fondé sur l'analyse markovienne de classes latentes (MCLCA), l'article de Biemer porte sur une évaluation de la convergence au cours du temps des réponses à la série de questions sur l'activité. L'accroissement de la non-

convergence entre la nouvelle série et l'ancienne, quand on tient compte de l'effet de l'auto-déclaration par opposition à la déclaration par procuration, peut servir d'indicateur d'un type d'erreur de mesure dans l'attribution de la situation d'activité. L'auteur présume que cette erreur est due au fait que les nouvelles questions (du moins, comparativement aux anciennes) ne permettent pas de recueillir l'information appropriée pour classer un individu dans la catégorie correcte de situation d'activité. Donc, selon lui, l'erreur peut être attribuée à une mauvaise conception du questionnaire. Comme, selon l'analyse, les erreurs ont tendance à aller dans une direction particulière, c'est-à-dire la classification erronée des personnes réellement en chômage dans une catégorie différente, d'aucuns pourraient interpréter le résultat comme un biais dans le taux de chômage.

Je soutiendrais non seulement qu'aucun biais n'a été introduit, mais aussi que la nouvelle série, quoique strictement imparfaite, réduit l'erreur et brosse un tableau plus exact de la situation d'activité, car elle tient compte de la conjoncture économique d'aujourd'hui d'une façon qui échappait à l'ancienne série. Cette amélioration est due non seulement à un meilleur énoncé des questions, mais aussi à l'intégration de questions de suivi et d'approfondissement qui fournissent des renseignements plus détaillés en vue de déterminer la situation d'activité réelle d'un répondant. L'introduction d'un questionnaire informatisé facilite l'utilisation de ces innovations, je suis convaincu que la nouvelle série de données sur l'activité réduit le nombre d'erreurs de spécification que comportait l'ancienne. Par erreur de

spécification, j'entends l'erreur qu'entraîne l'utilisation de questions qui ne mesurent pas ce qu'elles sont destinées à mesurer. J'expliquerais aussi pourquoi je ne pense pas que l'utilisation de la méthode de Biemer soit appropriée dans ce cas particulier.

2. RECONNAISSANCE DE LA NÉCESSITÉ D'UNE NOUVELLE SÉRIE DE DONNÉES SUR L'ACTIVITÉ

La dernière révision importante de la CPS, avant celle de 1994, a eu lieu en 1967. Durant les années qui ont suivi, le marché du travail a subi de grandes transformations. Le nombre de femmes faisant partie de la population active a augmenté spectaculairement. Les nombres d'emplois à temps partiel et de personnes détenant plusieurs emplois ont grimpé en flèche. La relation entre le travailleur et l'employeur est devenue plus ténue. Des progrès techniques étouffants ont modifié la façon de travailler des Américains et abouti à la création de nouvelles catégories d'emplois exigeant de nouvelles formes de compétences. Fait peut-être encore plus important, on a assisté progressivement à un essor du secteur des services et à un déclin du secteur de la fabrication.

L'une des conséquences de ces transformations dont il a fallu tenir compte dans la CPS est l'évolution du sens donné à l'expression « mise à pied » comme l'ont si expertement décrit Miller et Polivka (2004), mais aussi d'autres auteurs, comme le mentionnent Bregger et Dippo (1993). De meilleurs renseignements étaient nécessaires sur les travailleurs découragés (ceux qui décrochent et cessent de rechercher un emploi), les titulaires de plusieurs emplois, les travailleurs marginaux (par exemple, les travailleurs à temps partiel), les profils de changement d'emploi. En outre, au cours des années 1970 et 1980, on s'est soucié de plus en plus des divers types d'erreurs non dues à l'échantillonnage susceptibles d'affecter les estimations fondées sur la CPS, ainsi que du fardeau de réponse et de son effet indésirable sur la qualité des données. Jusqu'aux années 1980, la technologie permettant de s'attaquer à ces problèmes n'existait pas. Cependant, comme le font remarquer Bregger et Dippo (pages 4 et 5), les choses ont commencé à changer :

¹ Clyde Tucker, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Room 1950, Washington, D.C. 20212.

l'évolution de l'économie globale. Enfin, nous souhaitons que des travaux plus approfondis soient réalisés dans ce domaine en alliant la modélisation MILCA à une prise en compte minutieuse des concepts économiques mesurés, des périodes de référence examinées et des hypothèses formulées. Nous sommes convaincus que ce genre de travaux permettrait de mieux comprendre les effets du remaniement de la CPS de 1994 et d'appliquer plus efficacement l'approche de la modélisation MILCA en général.

REMERCIEMENTS

Les opinions exprimées dans le présent article sont celles des auteurs et ne constituent pas une politique du Bureau of Labor Statistics. Les auteurs remercient Sharon Cohany, du U.S. Bureau of Labor Statistics, de ses commentaires utiles sur la présente discussion.

BIBLIOGRAPHIE

DARBY, M.R., HALTIWANGER, J. et PLANT, M. (1985). Unemployment rate dynamics and persistent unemployment under rational expectations. *American Economic Review*, 75, 614-637.

GIBBONS, R., et KATZ, L.F. (1991). Layoffs and Lemons. *Journal of Labor Economics*, 9, 351-380.

MADRIAN, B.C., et LEFGREN, L.J. (1999). A Note on Longitudinally Matching Current Population Survey (CPS) Respondents. Document de travail technique, 247, *National Bureau of Economic Research Technical Working Paper Series*.

PALMISANO, M. (1989). Respondents' Understanding of Key Labor Force Concepts Used in the CPS. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria VA.

POLIVKA, A.E., et MILLER, S.M. (1998). The CPS After the Redesign: Refocusing the Economic Lens. Dans *Labor Statistics Measurement Issues*, (Eds., J. Haltiwanger, M.E. Manser et R. Topel). National Bureau of Economic Research Studies in Income and Wealth, Chicago: University of Chicago Press, 60, 249-286.

POLIVKA, A.E., et ROTHGEB, J. (1993). Overhauling the Current Population Survey: Redesigning the Questionnaire. *Monthly Labor Review*, 116, 10-28.

ROTHGEB, J. (1982). Summary Report of July Follow-up of the Unemployed. U.S. Bureau of the Census Memorandum, Washington D.C.

personne, cette hypothèse n'est manifestement pas vérifiée et sa validité diffère vraisemblablement pour les diverses catégories d'activité. Ainsi, une personne qui est occupée le premier mois est beaucoup plus susceptible d'être occupée le troisième mois qu'une personne qui n'a jamais travaillé. Point encore plus important, une personne ne peut être classée dans la catégorie des personnes mises à pied ni au moyen du questionnaire original ni au moyen du questionnaire révisé si elle n'a pas travaillé antérieurement. Cependant, aux termes de la définition officielle de la mise à pied qui a été appliquée dans le remaniement, les personnes doivent aussi s'attendre à être rappelées au travail. Cette condition donne lieu à une relation nettement plus étroite entre les employeurs et les travailleurs pour les mois où le questionnaire remanié a été utilisé. Étant donné que les personnes considérées comme mises à pied conformément au questionnaire révisé sont nettement plus susceptibles d'être rappelées au travail et, donc, occupées que celles classées ainsi aux termes du questionnaire non révisé, la probabilité que la situation d'activité d'une personne durant le troisième mois dépende de sa situation d'activité initiale durant le premier mois est nettement plus forte. Par conséquent, il est non seulement probable que les hypothèses markoviennes soient violées dans le cadre des études de la population active, mais il est aussi nettement plus probable qu'elles le soient après le remaniement. Cette violation différentielle des hypothèses du modèle pourrait influencer fondamentalement les résultats de l'auteur.

7. CONCLUSION

Sommairement, alors que l'auteur pense avoir identifié un problème introduit dans la CPS lors du remaniement de 1994, la prétendue augmentation de l'erreur de classification des personnes mises à pied reflète, en réalité, la plus grande précision des questions de l'enquête. Au lieu de repérer une erreur réelle, nous pensons que l'auteur pourrait avoir omis de reconnaître que la composition des groupes considérés comme étant en mise à pied avant et après le remaniement est différente à cause de modifications intentionnelles (comme l'intégration de la définition des personnes mises à pied dans le questionnaire et l'amélioration de la qualité des appariements obtenue grâce à l'informatisation de l'enquête) et à des changements non contrôlés, comme

Il est probable que la proportion de personnes qui demandent en chômage de moins en moins importante quand le taux de chômage est élevé que quand il est faible. À mesure que la conjoncture économique s'améliore et que le taux de chômage baisse, il n'est pas de raisonnable s'attendre à une augmentation de la proportion d'individus qui passent de la catégorie des personnes mises à pied à celle des personnes occupées. Parallèlement à l'augmentation du nombre de transitions vers l'emploi, la proportion de personnes qui passent à un état d'emploi temporaire pourrait également augmenter. En effet, bien qu'il soit indubitablement relié à de nombreux facteurs, le nombre de personnes occupées dans la classe de la location de personnel suppléant (telle qu'elle est définie dans le SCIAN) a augmenté de 44 % de 1992 à 1995, pour passer de 1,1 % à 1,5 % des listes de paye des établissements américains (mesuré par l'enquête auprès des établissements du BLS).

En outre, à mesure que baisse le taux de chômage, le genre de personnes classées comme étant en chômage peut changer. Plus précisément, celles qui demeurent en chômage quand le taux de chômage est faible ont tendance à trouver plus difficile d'obtenir un emploi et sont plus susceptibles de connaître des transitions rapides entre états d'activité. Ce raisonnement est celui qui sous-tend les études visant à analyser les effets de divers types de cessation d'emploi sur les résultats subséquents sur le marché du travail. Par exemple, lors d'une étude visant à comparer les personnes mises à pied par leur employeur à cause d'une conjoncture économique défavorable plutôt qu'à cause d'une fermeture complète d'usine, Gibbons et Katz (1991) observent qu'en ce qui concerne la durée de la période sans emploi et les gains, les travailleurs congédiés par leur employeur à cause d'une conjoncture économique défavorable se trouvaient dans une situation nettement plus défavorable que ceux ayant perdu leur emploi à cause d'une fermeture d'usine. Gibbons et Katz soutiennent que ces différences sont dues au fait que les employeurs peuvent congédier leurs travailleurs les moins productifs et retenir ceux qui sont les plus productifs quand la conjoncture économique est défavorable, alors que ceux dont l'établissement ferme complètement doivent congédier les travailleurs les plus productifs et les moins productifs. Parallèlement, Darby, Haitiwanger et Plant (1985) soutiennent que la durée du chômage augmente à mesure que la conjoncture économique se détériore, à cause d'un changement de composition du groupe de personnes en chômage. Cette situation est due au fait que, quand la conjoncture économique devient plus défavorable, la proportion de travailleurs hautement qualifiés en chômage (travailleurs qui sont aussi moins habitués à être chômeurs et plus susceptibles de pouvoir et de vouloir attendre de trouver un emploi plus

En plus des différences de composition du groupe de personnes classées comme étant mises à pied sur les estimations produites par les modèles MLCA, les différences de composition des groupes de personnes classées dans les diverses catégories d'activité avant et après le remaniement pourraient influencer sur la validité des hypothèses sous-jacentes des modèles MLCA. Comme le fait remarquer l'auteur, une hypothèse importante lors de la mise en œuvre de modèles MLCA est que la transition d'une personne du deuxième au troisième mois soit indépendante et ne soit donc pas influencée par le classement de cette personne durant le premier mois. Quand on estime des modèles MLCA pour les états d'activité d'une

6. VALIDITÉ DIFFÉRENTIELLE DES HYPOTHÈSES MARKOVIENNES

Il est important de souligner que la majorité des satisfaisants) augmentent et la proportion de travailleurs peu qualifiés, passent fréquemment d'un état d'activité à l'autre, qui sont en chômage diminue.

Il est important de souligner que la majorité des estimations produites par l'auteur pour la période avant le remaniement l'ont été au moyen de données recueillies pour 1992 et 1993, au moment où le taux de chômage moyen était de 7,0 %, tandis que la majorité des estimations pour la période après le remaniement ont été produites d'après des données recueillies pour 1994 et 1995, quand le taux de chômage moyen était de 6,0 %. L'évolution de la conjoncture économique générale et l'évolution correspondante de la composition du groupe de personnes en chômage pourraient influencer l'exactitude hypothétique des estimations de l'auteur d'une façon n'ayant aucun lien avec le questionnaire. Ainsi, de 1992 à 1995, la proportion d'adolescents en chômage a augmenté régulièrement pour passer de 14,8 % à 18,2 %, tandis que le taux global de chômage a baissé régulièrement, pour passer de 7,5 % à 5,6 %. De la même façon, la proportion d'Hispaniques en chômage a augmenté régulièrement pour passer de 13,6 % à 15,4 % de 1992 à 1995, bien qu'une part de cette augmentation puisse être attribuée à la proportion croissante d'Hispaniques dans la population (proportion qui est passée de 8,8 % à 9,4 %). Aussi bien les adolescents que les Hispaniques ont tendance à être des travailleurs peu qualifiés qui, traditionnellement, sont plus susceptibles que les autres de devenir chômeurs ou de se retirer du marché du travail. Il convient de souligner que, quelle que soit la source, l'augmentation de la proportion de chômeurs provenant de groupes caractérisés par des antécédents d'activité peu stables influencera les estimations de l'exactitude au moyen du modèle MLCA s'il n'est pas tenu compte du changement dans la modélisation.

À l'aide des mêmes données à grande diffusion que celles utilisées par l'auteur, conjuguées à des renseignements supplémentaires indiquant si une personne avait déménagé (renseignements qui sont recueillis périodiquement dans le cadre de la CPS), Madhian et Lefgren (1999) ont estimé que, selon la rigueur du critère d'appariement utilisé, de 64 à 87 % des appariements éliminés comme étant invalides étaient probablement des appariements légitimes. En outre, ces auteurs ont noté que la traction d'appariements invalides qui auraient probablement dus être retenus dans l'ensemble de données en prenant pour critère le fait qu'une personne avait déménagé a diminué considérablement de 1993 à 1996 (puisque Madhian et Lefgren ont utilisé des données à grande diffusion, ils n'ont pu étudier la validité des appariements obtenus pour 1994 et 1995, ni pour ceux obtenus pour 1995 et 1996, parce que la capacité d'apparier ces données a été supprimée afin d'assurer la protection des renseignements personnels fournis par les personnes concernées). Selon Madhian et Lefgren, l'accroissement du nombre d'appariements valides pour 1996 et les années suivantes est dû à des améliorations attribuables au remaniement (il convient de souligner que, bien qu'un puisse obtenir un meilleur appariement en utilisant les données internes du BLS et du Census Bureau pour lesquelles aucune information n'a été supprimée, la qualité d'un appariement en utilisant les données internes sera quand même affectée par la méthode de collecte des données. Donc, la qualité de l'appariement sera meilleure après le remaniement qu'avant celui-ci). L'étude de Madhian et Lefgren montre aussi qu'il était nettement plus probable que les personnes incorrectement exclues des ensembles de données appariées soient jeunes et que l'information à leur sujet ait été fournie par un autre membre du ménage (répondant par procuration). Ces personnes sont aussi celles qui, comme le soutient Biemer, sont plus susceptibles d'être classées dans une catégorie d'activité incorrecte. Par conséquent, il se pourrait qu'en incluant peut-être un plus grand nombre de ces personnes dans son étude à cause de la meilleure qualité de l'appariement, l'auteur ait observé une diminution de l'exactitude de ses mesures qu'il attribue incorrectement au questionnaire.

Une deuxième raison pour laquelle la composition des groupes correspondait aux divers états d'activité pourrait être différente pour les données recueillies selon la méthodologie originale et selon la méthodologie remaniée tient à la capacité d'établir la concordance entre les données individuelles de mois en mois et à la qualité de ces appariements. La grande majorité des données recueillies selon la méthodologie originale dans le cadre de la CPS officielle de 1993 ou dans celui de l'Enquête parallèle de 1994 a été au moyen d'un questionnaire papier-crayon et des intervieweurs ont dû transcrire manuellement les numéros d'identification du ménage et de la personne tirés des fichiers principaux sur les questionnaires imprimés. Toutes les données recueillies selon la méthodologie remaniée, dans le cadre de la CPS officielle après janvier 1994 ou dans celui de l'Enquête parallèle en 1993, l'ont été au moyen d'un ordinateur automatisé qui avait été téléchargé sur un ordinateur portable ou sur un ordinateur central. Dans le cadre du processus de collecte informatisé des données, les numéros d'identification du ménage et de la personne ont été automatiquement et uniformément reportés de mois en mois. L'utilisation de questionnaires imprimés et la transcription manuelle des données posent le risque d'introduire des erreurs et obligent les chercheurs à éliminer des enregistrements individuels non appariés qui correspondent en fait à une même personne et constituent donc des vrais appariements.

4. CHANGEMENTS TECHNOLOGIQUES POUR LA CUEILLETTE DES DONNÉES

Les deuxièmes raisons pour lesquelles la composition des groupes correspondait aux divers états d'activité pourraient être différentes pour les données recueillies selon la méthodologie originale et selon la méthodologie remaniée tiennent à la capacité d'établir la concordance entre les données individuelles de mois en mois et à la qualité de ces appariements. La grande majorité des données recueillies selon la méthodologie originale dans le cadre de la CPS officielle de 1993 ou dans celui de l'Enquête parallèle de 1994 a été au moyen d'un questionnaire papier-crayon et des intervieweurs ont dû transcrire manuellement les numéros d'identification du ménage et de la personne tirés des fichiers principaux sur les questionnaires imprimés. Toutes les données recueillies selon la méthodologie remaniée, dans le cadre de la CPS officielle après janvier 1994 ou dans celui de l'Enquête parallèle en 1993, l'ont été au moyen d'un ordinateur automatisé qui avait été téléchargé sur un ordinateur portable ou sur un ordinateur central. Dans le cadre du processus de collecte informatisé des données, les numéros d'identification du ménage et de la personne ont été automatiquement et uniformément reportés de mois en mois. L'utilisation de questionnaires imprimés et la transcription manuelle des données posent le risque d'introduire des erreurs et obligent les chercheurs à éliminer des enregistrements individuels non appariés qui correspondent en fait à une même personne et constituent donc des vrais appariements.

5. CONJONCTURE ÉCONOMIQUE

La conjoncture économique peut aussi faire varier la composition des groupes classés dans la catégorie des personnes mises à pied avant et après le remaniement. De 1992 à 1995, période que l'auteur utilise pour la majorité des ses modèles MLCA, le taux de chômage a diminué régulièrement. Plus précisément, en 1992, le taux annuel moyen de chômage était de 7,5 %, alors qu'en 1995, il était de 5,6 %.

La comparaison des estimations du modèle MLCa avant et après le remaniement sans tenir compte des différences de composition selon l'industrie des personnes classées dans la catégorie des mises à pied pourrait mener les analystes à conclure erronément que le remaniement a réduit l'exactitude de la classification de la situation d'activité. En réalité, l'augmentation du nombre de transitions mesurées après le remaniement représente une augmentation réelle du nombre de transitions vers l'emploi survenue après que les renseignements sur la mise à pied aient été demandés de façon appropriée au moyen du questionnaire de la CPS. Ne pas tenir compte du fait que le remaniement du questionnaire de la CPS donne lieu intentionnellement au classement dans la catégorie des personnes mises à pied d'un groupe de personnes légèrement différent de celui obtenu au moyen du questionnaire non révisé pourrait faire tirer des conclusions incorrectes des modèles MLCa. Les travailleurs ayant cessé définitivement de travailler pour leur employeur qui sont classés dans la catégorie des mises à pied au moyen du questionnaire non remanié semblent être plus correctement classés dans les modèles MLCa, mais ils sont en fait plus stables dans une classification incorrecte au départ. En outre, une certaine proportion de personnes correctement classées dans la catégorie des mises à pied selon la définition officielle pourraient, intrinsèquement, avoir des antécédents moins stables d'emploi à cause de leurs préférences personnelles ou des industries auxquelles elles sont associées.

À part les changements de composition liés aux différences dans l'énoncé des questions, l'auteur pourrait aussi avoir saisi par inadvertance dans ses estimations plusieurs autres changements de composition non liés à des différences d'énoncé. Ces changements incluent les différences concernant les périodes de référence utilisées par l'auteur pour produire ses estimations, ainsi que les modifications techniques apportées au processus de collecte de données et l'évolution de la conjoncture économique.

3. CARACTÈRE SAISONNIER

La première différence de composition que l'auteur pourrait avoir introduit par inadvertance est liée à la saisonnalité et aux différentes périodes de référence qu'il utilise pour l'estimation. Le nombre de personnes classées dans la catégorie des mises à pied par la CPS présente une importance variable saisonnière généralement caractérisée par un nombre plus élevé de personnes mises à pied au début de l'année. Par exemple, pour 358 personnes classées comme étant mises à pied en janvier 1995, il existait une concordance avec février et mars, pour 294 personnes concordance avec avril et mai et pour 188 personnes

est égal à l'effet de la nouvelle méthode moins l'effet de l'Enquête parallèle) indique que le remaniement a fait baisser la proportion de personnes classées dans la catégorie des mises à pied ayant anciennement un emploi dans le secteur de la fabrication de biens durables de 7,3 points et a fait augmenter la proportion de celles ayant anciennement un emploi dans le secteur de la construction de 3,7 points (le calcul de la moyenne de l'écart moyen entre le premier volet de l'Enquête parallèle et la CPS et de l'écart moyen entre la CPS et le deuxième volet de l'Enquête parallèle s'inscrit, quoique sous une forme simplifiée, dans la lignée des estimations des modèles linéaires des effets principaux par la méthode des moindres carrés généralisés présentée dans Polivka et Miller).

Les profils des transitions réelles sur le marché du travail des personnes appartenant à différentes industries pourraient être fort dissimilables, ce qui, à son tour, pourrait influencer les estimations MLCa. Par exemple, sachant qu'une proportion importante d'emplois du secteur de la construction sont sensibles aux conditions météorologiques et pourraient être plus axés sur les projets que d'autres types d'emplois, il n'est pas déraisonnable de s'attendre à ce que les travailleurs de la construction soient vraiment plus susceptibles d'être en mise à pied temporaire durant le premier des trois mois consécutifs, occupés à court terme durant le deuxième mois (parce que le temps s'est amélioré durant le deuxième mois ou qu'un projet de construction à court terme a été entrepris), puis de nouveau temporairement mis à pied durant le troisième mois (parce que les conditions météorologiques se sont détériorées ou que le projet pour lequel ils avaient été engagés s'est achevé). Par ailleurs, l'emploi dans le secteur de la fabrication de biens durables diminue régulièrement depuis les années 1970 (par exemple, en comparant les années sans récession, on a estimé qu'en 1971, 14,9 % de travailleurs américains, selon l'enquête auprès des établissements du BLS, étaient employés dans le secteur de la fabrication de biens durables comparativement à 9,2 % en 1993 et à 8,5 % en 2000). Étant donné ce recul de l'emploi, il est probable qu'une forte proportion de travailleurs du secteur de la fabrication classés dans la catégorie « mise à pied » avant le remaniement avaient cessé définitivement de travailler pour leur employeur (l'évolution de la répartition selon l'industrie observée après avoir imposé la contrainte de l'attente d'un rappel au travail corrobore cette notion). Le fait d'avoir perdu définitivement son emploi, conjugué à la fabrication de biens durables, pourrait accroître la probabilité que ces personnes soient en chômage pendant trois mois consécutifs, parce qu'il faut du temps pour trouver un emploi dans une autre industrie au même salaire.

Le remaniement aurait fait baisser la proportion d'hommes mis à pied et augmenter celle des femmes mises à pied comparativement aux proportions obtenues avant le remaniement, toutes choses étant égales par ailleurs. La comparaison des moyennes annuelles pour les personnes de plus de 20 ans appuie cette notion, puisqu'elle montre qu'en 1993, 67,2 % des personnes mises à pied étaient des hommes, comparativement à 63,6 % en 1994 (quoique, pour le remaniement du questionnaire, ces proportions pourraient être influencées par des changements de conjoncture économique).

La répartition selon l'industrie des personnes classées dans la catégorie des mises à pied au moyen des données des deux volets de l'Enquête parallèle et de la CPS officielle révèle d'autres changements de composition du groupe classé comme étant mis à pied avant et après le remaniement. La comparaison des estimations d'après l'enquête remaniée aux estimations officielles de la CPS pour janvier à mai 1993 et celle calculée d'après l'enquête non remaniée aux estimations officielles de la CPS pour janvier à mai 1994 révèle des écarts particulièrement spectaculaires pour l'industrie de la fabrication de biens durables. La proportion de personnes mises à pied qui étaient anciennement employées dans le secteur de la fabrication de biens durables au moment où les questions remaniées ont été utilisées est presque deux fois plus faible que celle obtenue quand on a utilisé le questionnaire remanié (pour janvier à mai 1993, la proportion de personnes mises à pied qui étaient anciennement employées dans le secteur des biens durables est, en moyenne, de 16,8 % pour celles qui ont répondu aux questions non remaniées et de 9,8 % pour celles qui ont répondu aux questions remaniées). Pour janvier à mai 1994, les proportions sont de 8,7 % pour celles qui ont répondu aux questions non remaniées et de 15,5 % pour celles qui ont répondu aux questions remaniées).

Parallèlement, la proportion de personnes mises à pied dans le secteur de la construction déterminée en utilisant les questions remaniées est supérieure de 10 à 15 % à celle qui avaient anciennement un emploi dans le secteur de la construction est de 33,3 %, en moyenne, pour celles qui ont répondu au questionnaire remanié et de 27,4 % pour celles qui ont répondu au questionnaire non remanié. Pour janvier à mai 1994, les proportions sont de 33,3 % et de 25,9 %, respectivement).

Le calcul de la moyenne de l'écart moyen entre le premier volet de l'Enquête parallèle et la CPS pour janvier 1993 à mai 1993 (écart qui correspond à l'effet de la nouvelle méthode plus l'effet de l'Enquête parallèle) et de l'écart moyen entre la CPS et le deuxième volet de l'Enquête parallèle pour janvier 1994 à mai 1994 (écart qui

des erreurs de classification des personnes mises à pied dans l'enquête remaniée sont dues aux questions sur l'attente d'un rappel au travail (« Votre employeur vous a-t-il donné une date de retour au travail? » et « Avez-vous des raisons de croire que vous serez rappelé(e) au travail d'ici six mois? »). Ce résultat cadre avec l'évolution des attentes de rappel et une légère augmentation du flux de l'état de personne mise à pied vers celui de personne inactive. L'auteur obtient des estimations MLCA différentes du taux de personnes considérées comme mises à pied avant et après le remaniement parce que la composition de ce groupe a été modifiée, et cette modification de la composition est celle qui était souhaitée et voulue par les auteurs du remaniement

D'autres preuves du changement de composition du questionnaire. L'adoption du nouveau questionnaire en 1994, l'ancien questionnaire a été administré mensuellement de janvier 1994 à mai 1994 à 12 000 ménages tirés à partir du même échantillon. L'utilisation expérimentale des questionnaires originaux et remaniés a reçu le nom d'« Enquête parallèle ». Les estimations d'après l'Enquête parallèle réalisées avant 1994 en utilisant la nouvelle méthodologie et calculées en vue de leur comparaison aux estimations officielles de la CPS basées sur les procédures remaniées de la CPS avant 1994 et sur les procédures remaniées après 1994, Polivka et Miller (1998) illustrent l'importance qu'il y a à utiliser les deux volets de l'Enquête parallèle pour obtenir un tableau complet des effets du remaniement de l'enquête. Par exemple, si l'on utilise uniquement le premier volet de l'Enquête parallèle, on estime que le remaniement a augmenté le taux de chômage de 0,5 point. Par contre, si l'on utilise les deux volets de cette enquête, les estimations indiquent que le remaniement n'a eu aucun effet statistiquement significatif sur le taux de chômage.

En utilisant les deux volets de l'Enquête parallèle et les estimations officielles de la CPS, Polivka et Miller estiment que le remaniement de la CPS a produit une baisse d'un peu moins de 7 % de la proportion de chômeurs de sexe masculin en mise à pied et qu'elle a fait augmenter de presque 7 % (quoique cette dernière estimation ne soit pas statistiquement significative au niveau de signification de 5 %) la proportion des chômeuses classées dans la catégorie des personnes mises à pied. Ces estimations impliquent que

classification (ces dernières peuvent être classées dans la catégorie des personnes en chômage à une étape ultérieure du questionnaire si elles satisfont aux critères de recherche active d'un emploi et de disponibilité).

À cause de l'ajout de ces questions directes, on devrait s'attendre à ce que le groupe de personnes classées dans la catégorie des mises à pied soit légèrement différent. Avant le remaniement, une proportion importante, voire la majorité, des personnes classées comme étant mises à pied avaient, en fait, cessé définitivement de travailler pour leur employeur. Après le remaniement, les personnes considérées comme étant mises à pied devaient s'attendre à être rappelées au travail par leur ancien employeur, de sorte que la grande majorité d'entre elles devaient n'avoir cessé de travailler pour leur employeur que temporairement. Il n'est pas étonnant que ces deux groupes de personnes soient caractérisés par des flux de mois en mois différents entre les catégories d'activité. Il est raisonnable de s'attendre à ce que les personnes pensant qu'elles seront rappelées au travail soient plus susceptibles de passer durant des mois consécutifs de l'état de mise à pied temporaire à l'état de personne occupée que celles ayant perdu définitivement leur emploi. En outre, comparativement aux travailleurs ayant perdu définitivement leur emploi, ceux appartenant aux industries seraient plus susceptibles d'être mis à pied un mois, occupés le mois suivant, puis de nouveau mis à pied.

Les flux bruts de mois en mois entre les états d'activité indiquent que la proportion de personnes qui sont passées du chômage à un emploi a augmenté après le remaniement de 1994. Plus précisément, en 1994, 26,6 % des personnes en chômage durant le premier mois étaient occupées durant le second, comparativement à 23,7 % en 1993.

Les estimations MLC A calculées par l'auteur de la diminution supposée de l'exactitude du classement dans la catégorie des personnes en mise à pied après le remaniement à cause du classement d'un plus grand nombre de personnes dans la catégorie des personnes occupées après qu'elles aient été mises à pied concordent, en fait, exactement avec les résultats on s'attendrait à la suite d'un resserrement de la définition de la mise à pied, et est en accord avec l'augmentation des flux bruts de mois en mois entre les états de chômage et d'emploi (quoique l'augmentation du flux cadre aussi avec la baisse du taux de chômage observée durant la période couverte par l'étude de l'auteur). L'estimation MLC A plus faible, mais néanmoins significative, de la diminution de l'exactitude due au fait qu'un plus grand nombre de personnes mises à pied ont été considérées comme étant inactives après le remaniement concorde aussi avec le resserrement de la définition de la mise à pied résultant de l'exigence que ces personnes s'attendent à être rappelées au travail dans les six mois qui

suivent, puisque certaines personnes pourraient s'adapter ou modifier leurs attentes de rappel au travail au cours du temps. Par exemple, au moment de la première interview, ces personnes pourraient s'attendre à être rappelées dans les six mois qui suivent. Toutefois, les mois suivants, à mesure que le temps écoulé depuis la cessation d'emploi initiale s'accroît, ces personnes pourraient ne plus déclarer qu'elles s'attendent à être rappelées au travail. Si, parallèlement, elles n'ont pas commencé à rechercher un autre emploi, peut-être parce qu'elles ont encore droit aux prestations d'assurance-chômage, ces personnes pourraient passer à l'état de personnes inactives. Ou bien, certaines personnes pourraient croire qu'elles vont être rappelées au travail, mais, au cours des mois suivants, à cause d'intempéries ou d'une détérioration de la conjoncture économique de leur ancien employeur, devenir moins certaines d'un rappel au travail et donc ne plus déclarer qu'elles s'attendent à être rappelées. Si, au cours des mois suivants, la conjoncture économique de leur ancien employeur s'améliore ou que le temps devient plus clément, ces personnes pourraient de nouveau penser correctement qu'elles seront rappelées au travail. L'évolution des attentes pourrait produire un profil de trois mois où certaines personnes étaient réellement mises à pied le premier mois, inactives le deuxième mois, et de nouvelles mises à pied le troisième mois. Celles dont l'emploi s'est terminé définitivement et qui ont été classées incorrectement dans la catégorie des personnes mises à pied avant le remaniement de l'enquête ne seraient pas touchées par l'évolution des attentes de rappel. Par conséquent, les personnes dont l'emploi s'est terminé définitivement seraient probablement plus susceptibles de se déclarer elles-mêmes en mise à pied durant les mois consécutifs dans le cas de l'enquête non remaniée. Le modèle MLC A interpréterait cette plus grande stabilité comme une indication que la mesure du taux de mises à pied était plus précise avant le remaniement. Cependant, cette plus grande « exactitude » ne s'appliquerait qu'aux personnes classées incorrectement parce qu'elles ont utilisé une définition trop générale.

L'auteur conclut que 60 % des cas de classification erronée des personnes mises à pied observés pour l'enquête remaniée sont dus à la question « La SEMAINE DERNIÈRE, avez-vous fait DU travail contre rémunération? ». En réalité, ce résultat est conforme au fait qu'un plus grand nombre de personnes sont considérées comme temporairement mises à pied et rappelées au travail par leur ancien employeur dans l'enquête remaniée (quoique, si les personnes mises à pied prennent un emploi temporaire en entendant d'être rappelées par leur ancien employeur, une augmentation des transitions vers l'emploi après 1994 pourrait aussi être due, du moins partiellement, à la question plus générale sur l'emploi utilisée dans l'enquête

Commentaire

STEPHEN M. MILLER et ANNE E. POLIVKA¹

1. INTRODUCTION

2. AMÉLIORATION DE LA MESURE

L'une des raisons principales du remaniement de la CPS était de mesurer plus exactement les définitions et les concepts officiels. Le concept de la mise à pied était particulièrement problématique, en ce sens que la signification qui lui était généralement attribuée durant les années 1990, c'est-à-dire une cessation d'emploi permanente, était fort différente de la définition officielle de la CPS, c'est-à-dire une cessation d'emploi temporaire avec l'attente d'un rappel au travail. Au moment de la rédaction originale des questions durant les années 1940, l'expression mise à pied était utilisée couramment pour désigner des périodes de chômage temporaires dues à un roullage ou à une détérioration de la conjoncture économique. Par conséquent, le questionnaire utilisé avant le remaniement ne contenait pas de questions sur les attentes de rappel au travail. Des travaux de recherche réalisés durant les années 1980 et au début des années 1990 en prévision du remaniement de l'enquête ont indiqué que l'interprétation de l'expression mise à pied par les répondants était devenue considérablement plus générale que la définition officielle. La conduite de groupes de discussion et un questionnaire approfondi à grande échelle des répondants ont montré que de 30 à 50 % des personnes qui déclareraient avoir été mises à pied ne s'attendaient pas à retourner travailler pour leur ancien employeur (Rothgeb 1982; Palmisano 1989; Polivka et Rothgeb 1993). En outre, en 1993, 5,4 % de personnes classées dans la catégorie des mises à pied avaient travaillé pour la dernière fois d'une à cinq années plus tôt et 0,6 % n'avaient pas travaillé au cours des cinq dernières années. Ce manque d'antécédents de travail récents appuie la thèse voulant que nombre de personnes classées dans la catégorie des mises à pied avant le remaniement n'avaient aucun espoir d'être rappelées au travail.

Pour mieux mesurer la définition officielle de la mise à pied de la CPS, deux questions portant sur les attentes de rappel ont été ajoutées au questionnaire révisé – « Votre employeur vous a-t-il donné une date de retour au travail? » et « Avez-vous des raisons de croire que vous serez rappelé(e) au travail d'ici six mois? ». Les personnes qui répondaient affirmativement à l'une de ces questions sont classées dans la catégorie des mises à pied si elles sont capables de travailler; les autres sont exclues de cette

Nous sommes reconnaissants d'avoir l'occasion de commenter cet article intéressant. Nous commenterons principalement les observations empiriques au sujet du remaniement de la Current Population Survey (CPS) de 1994, plutôt que de nous lancer dans une discussion technique de la méthodologie de l'analyse markovienne de classes latentes (MLCA) proprement dite.

Dans son article intitulé « Une analyse de l'erreur de classification pour les questions sur l'emploi révisées de la Current Population Survey », l'auteur se sert de modèles MLCA pour essayer de déceler la source de ce qu'il croit être la « réduction de l'exactitude de la classification révisée des personnes en chômage » après le remaniement. Dans la CPS, on considère comme étant en chômage les personnes qui sont classées dans la catégorie des personnes mises à pied ou dans celle des personnes à la recherche d'emploi. L'auteur fait part d'une réduction particulièrement importante de l'exactitude du taux de personnes mises à pied. Par conséquent, nous nous concentrerons sur la classification de ces personnes, quoique des commentaires semblables pourraient être faits au sujet de la variation de la mesure du taux de personnes à la recherche d'emploi. Pour examiner l'exactitude de la mesure du taux de personnes mises à pied, l'auteur suppose que les personnes classées dans cette catégorie sont conceptuellement les mêmes avant et après le remaniement de 1994 et que ces personnes devraient être caractérisées de mois en mois par des flux identiques sur le marché du travail. Pourtant, il existe de nombreuses raisons pour lesquelles les mesures améliorées intégrées dans le remaniement devraient modifier conceptuellement le classement des personnes mises à pied. En outre, plusieurs facteurs non corrélés à la modification de l'énoncé des questions sont susceptibles d'influencer la composition du groupe de personnes classées dans la catégorie des mises à pied. Par conséquent, ce que l'auteur décrit comme étant une réduction de l'exactitude due au remaniement pourrait être attribué plus correctement à une modification conceptuelle de la catégorie des personnes mises à pied et au fait que ce que mesurait la CPS avant le remaniement diffère de ce qu'elle mesure depuis le remaniement.

¹ Stephen M. Miller et Anne E. Polivka, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Room 4985, Washington, D.C. 20012.

comportant quatre catégories de situation d'activité, on pourrait utiliser l'ensemble complet de questions sur la situation d'activité de la CPS auxquelles ont répondu les participants à l'enquête. Ce genre d'analyse avec indicateurs multiples serait non seulement beaucoup plus informatif, mais permettrait aussi de vérifier et de relâcher certaines hypothèses émises au cours de l'analyse. Ainsi, l'hypothèse d'IEC pourrait être relâchée pour certains éléments du questionnaire.

BIBLIOGRAPHIE

- BASSI, F., HAGENARS, J.A., CROON, M. et VERMUNT, J.K. (2000). Estimating true changes when categorical panel data are affected by uncorrelated and correlated classification errors. *Sociological Methods and Research*, 29, 230-268.
- PAAS, L.J., BUMOLT, T.H. et VERMUNT, J.K. (2003). Extending dynamic Segmentation with Lead Generation: A Latent Class Markov Approach. Center Paper, Tilburg University (soumis pour publication).
- VERMUNT, J.K. (1997). *Log-linear models for event histories*. Techniques in the Social Sciences Series, Thousand Oaks: Sage Publications, 8.
- VERMUNT, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, Sous presse.
- VERMUNT, J.K., LANGHEINE, R. et BÖCKENHOLT, U. (1999). Latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24, 178-205.
- VERMUNT, J.K., et MAGIDSON, J. (2000). *Latent GOLD User's Manual*. Boston: Statistical Innovations Inc.
- VERMUNT, J.K., et MAGIDSON, J. (2003). *Addendum to Latent GOLD User's Guide: Upgrade for Version 3.0*. Boston: Statistical Innovations Inc.
- VERMUNT, J.K., RODRIGO, M.F. et ATO-GARCIA, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods and Research*, 30, 170-196.

Autrement dit, s'il existe, au niveau manifeste, 2^K au lieu de K réponses possibles. Même dans ce cas, il suffit souvent de conceptualiser le modèle sous forme d'un modèle à une variable latente avec K + 1 classes et K indicateurs, structure qui est parfois qualifiée de modèle probabiliste de Guttman.

Paul Biemer reconnaît la complexité de la formulation des K variables latentes et des K variables manifestes et choisit de simplifier le modèle. Cependant, à cause de ce point de départ je suppose, il a décidé de garder K + 1 classes latentes. Je ne vois pas pourquoi un si grand nombre de classes latentes est nécessaire. Il n'existe même pas autant de situations d'activité. Il serait plus logique de n'avoir que deux classes – « mise à pied » et « pas de mise à pied » (« recherche d'emploi » et « pas de recherche d'emploi ») – puisque les questions visent uniquement à faire cette distinction particulière. Il se pourrait naturellement que les questions fournissent des renseignements sur le type de situation « pas de mise à pied » (« pas de recherche d'emploi »), auquel cas une classe latente supplémentaire pourrait être nécessaire. Ce qui me paraît clair est que K + 1 constitue un beaucoup trop grand nombre de classes.

Je me demande combien de personnes ont été classées dans la catégorie « mise à pied » (« à la recherche d'emploi ») aux diverses périodes de l'analyse réalisée avec la variable composite R comme indicateur. Ces nombres, ainsi que les nombres de transitions vers cet état et hors de celui-ci sont-ils semblables à ceux obtenus au moyen du modèle markovien de classes latentes à quatre états standard. Selon moi, il s'agit d'une condition nécessaire pour que les calculs faits pour obtenir les chiffres présentés aux tableaux 3 et 4 soient valides.

Le dernier commentaire qui me vient à l'esprit est le suivant. Pourquoi ne pas construire un modèle markovien de classes latentes en utilisant toute l'information du questionnaire, comme dans la deuxième partie de l'analyse? Autrement dit, au lieu de la classification construite observée

période de trois mois pendant laquelle un groupe de renouvellement est interviewé, mais qu'elles peuvent varier d'un groupe de renouvellement à l'autre, même si ceux-ci sont interviewés durant le même mois. Il serait naturel-ment préférable d'imposer des contraintes d'égalité sur les groupes de renouvellement. Une application uniforme de l'hypothèse d'homogénéité dans le temps impliquerait que, tant pour le questionnaire original que pour le questionnaire révisé, les erreurs de mesure soient constantes durant la période d'étude complète.

Ce dont nous avons effectivement besoin est un modèle markovien de classes latentes couvrant l'ensemble des 30 mois; autrement dit, un modèle pour 30 points dans le temps au lieu de 3. La spécification d'un modèle simultané de ce genre pour l'ensemble des groupes de renouvellement est aussi aisée que celle d'un modèle pour trois points dans le temps. Évidemment, pour chaque groupe de renouvellement, on n'observe que trois des 30 mois, si bien que les autres points dans le temps doivent être traités comme des valeurs manquantes. Cela ne pose aucun problème pour l'estimation du maximum de vraisemblance des paramètres du modèle, puisque nous pouvons simplement supposer que les données manquent au hasard (Vermunt 1997). Le type de questionnaire (original/révisé) sert de variable de groupe-ment (en plus du mode d'interview) et influe sur les probabilités d'erreur de classification homogène dans le temps. Autrement dit, nous estimons uniquement deux ensembles d'erreurs de classification, l'un pour le question-naire original et l'autre pour le questionnaire révisé. Les probabilités de transition peuvent varier en fonction du temps, mais seront égales pour tous les groupes de renouvellement interviewés durant un même cycle. En outre, les probabilités initiales des états pour un groupe de renouvellement ne sont pas estimées comme des paramètres distincts, puisqu'elles sont définies par l'état courant de la chaîne de Markov latente.

Un problème pratique de la modélisation simultanée est qu'avant un aussi grand nombre de points dans le temps, il n'est plus possible d'estimer les paramètres du modèle au moyen de l'algorithme EM ordinaire. Cependant, au moyen d'une variante de cet algorithme, appelée algorithme de Baum-Welch, on peut appliquer le modèle à un grand nombre de périodes (Vermunt 2003; Paas, Bijmolt et Vermunt 2003). Cet algorithme est implémenté dans une version expérimentale du programme Latent GOLD (Vermunt et Magidson 2000, 2003) et sera disponible dans une prochaine version de ce programme.

Un autre moyen de mettre en œuvre un modèle simultané consiste à utiliser un modèle markovien de classes latentes couvrant trois cycles d'enquête dans lequel le groupe de renouvellement sert de variable de groupement et les contraintes d'égalité entre groupes de renouvellement

4. MODÈLE DU PROCESSUS DE RÉPONSE

Essayer de déterminer quelles questions du questionnaire causent les erreurs de classification par modélisation du processus de réponse proprement dit est une excellente idée. Cette analyse pourrait fournir une foule de renseignements précieux pour le remaniement du questionnaire. Néanmoins, j'estime personnellement que la formulation des modèles étendus proposés pour les situations d'activité « mise à pied » et « recherche d'emploi » est trop compliquée.

La forme de la variable créée R est la même que celle de la variable de résultat dans un choix séquentiel ou dans une analyse de survie en temps discret. Le fait que la réponse à la question courante soit affirmative ou non détermine entièrement la réponse à la question suivante. L'information dont nous disposons est le nombre d'étapes que parcourt une personne, ce qui équivaut conceptuellement à une période de survie discrète. Une personne qui « survit » jusqu'au bout est classée dans la catégorie « mise à pied » (« recherche d'emploi »).

Selon moi, il n'est pas très utile de traiter cette variable comme étant générée par K variables latentes (les T). Cela n'est sensé que s'il devait théoriquement exister une hiérarchie de réponses au niveau latent, ce que l'erreur de mesure empêche toutefois d'observer au niveau manifeste.

D'autres aspects de la modélisation qui pourraient être perfectionnés sont le traitement des valeurs manquantes et le codage du mode d'interview. Il n'est pas nécessaire d'éliminer de l'analyse les cas pour lesquels des valeurs manquent, comme le fait Paul Biemer, parce que l'estima-tion du maximum de vraisemblance avec des valeurs manquantes est simple. En ce qui concerne le mode d'interview, il serait nettement plus élégant de ne travailler qu'avec deux catégories – par procuration et sans procuration – au lieu de quatre et de laisser le mode d'interview varier en fonction du cycle dans les cas. Autrement dit, le mode d'interview pourrait servir de covariable variable en fonction du temps. Vermunt, Langeheine et Böckenholt (1999) ont proposé un modèle markovien de classes latentes de ce genre avec covariables variables en fonction du temps.

Essayer de déterminer quelles questions du questionnaire causent les erreurs de classification par modélisation du processus de réponse proprement dit est une excellente idée. Cette analyse pourrait fournir une foule de renseignements précieux pour le remaniement du questionnaire. Néanmoins, j'estime personnellement que la formulation des modèles étendus proposés pour les situations d'activité « mise à pied » et « recherche d'emploi » est trop compliquée.

La forme de la variable créée R est la même que celle de la variable de résultat dans un choix séquentiel ou dans une analyse de survie en temps discret. Le fait que la réponse à la question courante soit affirmative ou non détermine entièrement la réponse à la question suivante. L'information dont nous disposons est le nombre d'étapes que parcourt une personne, ce qui équivaut conceptuellement à une période de survie discrète. Une personne qui « survit » jusqu'au bout est classée dans la catégorie « mise à pied » (« recherche d'emploi »).

Selon moi, il n'est pas très utile de traiter cette variable comme étant générée par K variables latentes (les T). Cela n'est sensé que s'il devait théoriquement exister une hiérarchie de réponses au niveau latent, ce que l'erreur de mesure empêche toutefois d'observer au niveau manifeste.

particulier dans le temps ou de la situation qui serait enregistré au moyen d'un instrument exempt d'erreur ou représentant l'étalon d'or? Ou s'agit-il de la situation qu'aurait occupée une personne dans des « conditions normales »? C'est-à-dire si l'on élimine aussi par filtration le caractère aléatoire de son comportement.

Voici un petit exemple pour illustrer ce point. Supposons qu'il existe deux types (deux segments latents) de consommateurs de café, à savoir ceux qui préfèrent la marque A et ceux qui préfèrent la marque B, et que j'appartienne au segment de la marque B, ce qui signifie que, dans des

circonstances normales, j'achète le café de marque B. Lors d'une interview, on me demande quelle marque j'ai achetée la semaine dernière. Supposons que je déclare avoir acheté un paquet de café de marque A et que je ne mente pas ni ne commette d'erreur. Autrement dit, il n'y a aucune erreur de classification au sens d'avoir fait une erreur : j'ai réellement acheté la marque A cette semaine-là (naturellement le chercheur ne le sait pas). Par contre, mon comportement cette semaine-là ne concorde pas avec mes préférences, ce qui signifie qu'en ce qui concerne la mesure de mes préférences, il y a une erreur de classification. Cet exemple illustre le fait que deux types d'erreur « peuvent être commis. Une erreur de déclaration et une « erreur » de comportement (l'« erreur » de comportement de la semaine précédente peut avoir de nombreuses causes, dont « l'épuisement du stock de la marque B », « la promotion de la marque A à un prix plus faible cette semaine-là », l'incapacité de trouver la marque B à cause d'une réorganisation des rayons du supermarché », etc. Le modèle markovien de classes latentes ne permet pas de faire la distinction entre ce caractère aléatoire, sans corrélation temporelle, du comportement et les erreurs réelles de classification.

Quelle est l'implication en ce qui concerne l'évaluation de la situation d'activité? Cela signifie que l'état réel d'un individu pourrait être « mise à pied », mais que, pour une raison particulière (par chance), ce mois-là, il ou elle a travaillé. Si cette « raison particulière » n'est pas contrôlée à d'autres « raisons particulières », le modèle « incorrect » durant d'autres cycles d'enquête, le modèle markovien de classes latentes indiquera qu'il s'agit d'une erreur de classification. Si, dans le cas de la mesure des préférences fondées sur les préférences révélées (ou déclarées), corriger pour le caractère aléatoire du comportement semble être exactement ce que l'on souhaite faire, il n'en est manifestement pas ainsi dans la mesure de la situation d'activité. Par conséquent, j'ai le sentiment que les taux d'erreur présentés par Biemer pourraient être quelque peu surestimés à cause du caractère aléatoire du comportement lié à l'activité d, par exemple au caractère aléatoire du fonctionnement du marché du travail.

Une conséquence bien connue de la modélisation des changements individuels au moyen d'un modèle markovien de classes latentes est que le nombre estimé de transitions latentes est nettement plus faible que le nombre observé correspondant. Il en est ainsi parce que les erreurs de classification indépendantes sont filtrés; autrement dit, une partie du changement observé est attribuée à ces phénomènes.

3. MODÈLE MARKOVIEEN DE CLASSES LATENTES: SPÉCIFICATION

Paul Biemer a estimé un modèle markovien de classes latentes couvrant trois cycles distincts pour chacun des 30 ensembles de données de trois mois. Il a utilisé le mode d'interview comme variable de groupement afin de tenir compte d'une partie de l'hétérogénéité des distributions réelles des situations d'activité et des erreurs de classification. Les taux d'erreur présentés dans les tableaux de des moyennes sur les modes d'interview et les groupes de renouvellement.

J'aurais personnellement spécifié le modèle d'une façon un peu plus élégante et moins arbitraire. Au lieu d'exécuter une analyse distincte pour chaque groupe de renouvellement, j'aurais essayé de construire un modèle simultané pour tous les groupes. Le problème majeur que pose l'exécution d'une série d'analyses distinctes est que les paramètres qui devraient effectivement être égaux pour tous les groupes de renouvellement sont, dans ces conditions, estimés sans contraintes. Ainsi, la répartition de l'activité en mars 1994 devrait être la même pour les groupes de renouvellement qui ont été interviewés de janvier à mars, de février à avril et de mars à mai, respectivement. En outre, les probabilités de transition (changement de situation d'activité) entre mars et avril devraient être les mêmes dans les groupes de renouvellement qui ont été interviewés de février à avril et de mars à mai. Ce fait a également des implications pour les groupes de l'Enquête parallèle : on devrait supposer que leurs distributions et transitions latentes selon la période correspondent à celles de la CPS ordinaire. Il se serait agité d'un moyen nettement meilleur de déterminer s'il existe une différence d'erreur de mesure entre les deux questionnaires. Surtout pour la période durant laquelle il y a chevauchement des deux formes de questionnaires, il est essentiel de supposer que les distributions latentes sont les mêmes afin d'empêcher que les différences entre les erreurs de mesure ressemblent partiellement à des différences entre états réels.

Les analyses distinctes posent un problème semblable pour l'estimation des erreurs de classification. Il est supposé que ces erreurs ne varient pas en fonction du temps durant la

Commentaire

JEROEN K. VERMUNT¹

1. INTRODUCTION

J'ai eu beaucoup de plaisir à lire cet article très bien écrit.

Le sujet abordé par Paul Biemer – les erreurs de classification dans la mesure de la situation d'activité – est très important. Les statistiques sur l'emploi font partie des indicateurs macroéconomiques les plus importants et, en fait, nous souhaiiterions qu'elles soient exemptes d'erreurs. Malheureusement, il s'avère impossible de mesurer l'activité d'une personne sans erreur. Le mieux qu'on puisse faire est concevoir la collecte des données de telle façon que les erreurs de classification au niveau individuel soient, dans la mesure du possible, réduites au minimum. L'article contribue à la réalisation de cet objectif.

Une étude antérieure de Biemer et Bushery (2000) a indiqué que les changements apportés en 1993 à la procédure de mesure dans le but de réduire les erreurs de classification avaient, en fait, accru l'erreur de mesure. Dans l'article courant, Paul Biemer répète les analyses antérieures sur une série chronologique plus longue, en ajoutant une catégorie d'activité supplémentaire obtenue par subdivision de la catégorie des personnes en chômage en une catégorie de personnes en « mise à pied » et une catégorie de personnes à la « recherche d'emploi ». Les résultats qu'il présente confirment la conclusion antérieure selon laquelle la nouvelle procédure donne de moins bons résultats que l'ancienne. En second lieu, Biemer essaye de décrire les sources de l'erreur de mesure pour les deux catégories de chômage en modélisant les questions distinctes utilisées pour déterminer si une personne est « mise à pied » ou « à la recherche d'emploi ». Il repère ainsi des sources d'erreur qui permettent d'envisager d'éventuelles améliorations du questionnaire.

Étant donné mon domaine de spécialisation, mon commentaire portera principalement sur les aspects méthodologiques et statistiques. Plus précisément, je discuterai de certains problèmes méthodologiques que posent l'application du modèle markovien de classes latentes et indiquerai comment l'analyse statistique pourrait être perfectionnée. Il n'est toutefois pas certain que cette modélisation plus élégante mène à des conclusions fort différentes. De nouveau, j'insiste sur le fait qu'il s'agit d'un excellent article. Mes critiques ont pour seul objectif d'amorcer la discussion.

2. MODÈLE MARKOVEN DE CLASSES LATENTES : MÉTHODOLOGIE

Les classes latentes du modèle de Markov caché sont le moteur principal de l'étude réalisée par Paul Biemer. Plusieurs hypothèses susceptibles d'influer sur les résultats observés doivent être éliminées, comme cela est le cas dans la présente étude, quand on applique le modèle avec un seul indicateur par cycle d'enquête. L'hypothèse discutée en détail par Biemer est celle du processus markovien d'ordre 1. Les études en simulation de Biemer et Bushery ont montré qu'heureusement, les estimations de l'erreur de classification ne sont pas très sensibles à cette hypothèse. Une autre hypothèse nécessaire ici pour l'identification du modèle est que l'erreur de mesure est constante au cours du temps. Cette hypothèse ne semble pas très problématique dans l'étude courante, puisque nous recherchons une mesure unique invariante dans le temps de l'erreur de classification. De surcroît, il n'existe aucune bonne raison de supposer que la qualité de la méthode de mesure a évolué au cours du temps alors que la méthode proprement dite n'a pas changé (à part, naturellement, le remaniement du questionnaire). Je suis nettement plus préoccupé par la troisième hypothèse, c'est-à-dire celle de l'indépendance des erreurs de classification (IEC) au cours du temps (Bassi, Hagenaars, Croon et Vermunt 2000). Est-il raisonnable de supposer que la survenue d'un certain type d'erreur de classification au point t dans le temps n'influence pas la probabilité de commettre la même erreur au point $t + 1$? À mon avis, cette hypothèse n'est pas judicieuse dans l'application qui nous préoccupe. Par exemple, un répondant qui commet une erreur parce qu'il ou elle n'a pas compris l'une des questions commettra fort probablement (ou du moins, sera plus susceptible que d'autres de le faire) de nouveau la même erreur lors du cycle suivant. Selon moi, il est nécessaire de réaliser une étude en simulation pour déterminer la sensibilité des estimations des erreurs de classification aux violations de l'hypothèse d'IEC.

J'ai une autre critique à formuler quant à l'utilisation du modèle markovien de classes latentes pour quantifier l'erreur de mesure de la situation d'activité d'une personne. Suivant le modèle, il existe une relation probabiliste entre les états réel et observé d'un individu. Cependant, quel est l'état réel? S'agit-il de la situation réelle d'activité à un point

Pour classifier les personnes à la recherche d'un emploi au moyen du questionnaire révisé, deux composantes du questionnaire semblent contribuer le plus à l'erreur de classification : « La SEMAINE DERNIERE, avez-vous fait TOUT travail (soit) contre rémunération (soit en vue d'un bénéfice?) » et « Avez-vous fait quoi que ce soit pour trouver du travail au cours des quatre dernières semaines?/Qu'avez-vous fait pour trouver du travail au cours des quatre dernières semaines? ». Pour les deux questions, le taux d'erreur est légèrement plus élevé dans le cas du questionnaire révisé que dans celui du questionnaire original. Par conséquent, ces augmentations expliquent la légère hausse du taux d'erreur de classification des personnes à la recherche d'un emploi au moyen du questionnaire révisé.

L'erreur de classification des personnes en chômage dans le cadre de la CPS est bien décrite; consulter, par exemple, Chua et Fuller 1987; Abowd et Zellner 1985; Portetba et Summers 1995; et Sinclair et Gastwirth 1998. Une mesure généralement reconnue de la fiabilité des données de la CPS, c'est-à-dire l'indice d'incohérence calculé d'après les données de réinterview de la CPS, indique que la fiabilité de la CPS a diminué après le remaniement de l'enquête. Les résultats présentés ici corroborent ceux des études antérieures et facilitent la détermination de la source de l'erreur de classification des personnes en chômage dans la CPS. À tout le moins, nos résultats serviront de tremplin pour d'autres études des causes profondes des erreurs qui entachent la collecte des données sur la population active au moyen de la CPS. Grâce à des expériences cognitives en laboratoire et à des essais sur le terrain, nous pourrions cerner les causes de l'erreur dans les questions sur le chômage, ce qui permettrait de dégager des moyens d'améliorer ces questions. Ce genre d'améliorations pourraient être mises en œuvre lors d'un futur remaniement de la CPS.

REMERCIEMENTS

L'auteur remercie de son aide Pamela McGovern, du U.S. Census Bureau, qui a commenté les premières ébauches de l'article. Il tient aussi à remercier le rédacteur associé et un examinateur anonyme dont les commentaires ont été fort utiles pour la préparation de l'article. L'étude a été financée par le U.S. Census Bureau.

BIBLIOGRAPHIE

ABOWD, J. et ZELLNER, A. (1985). Estimating gross Labor-Force Flows. *Journal of Business and Economic Statistics*, 3, 3, 254-283.

- BIEMER, P. (2004). Modeling measurement error to identify flawed questions. In *Methods for Testing and Evaluating Survey Questionnaires*. (Eds. S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, et E. Singer). Hoboken, New Jersey: John Wiley & Sons, Inc., 225-246.
- BIEMER, P., et BUSHERY, J. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. *Techniques d'enquête*, 26, 139-152.
- BIEMER, P., et WIESEN, C. (2002). Latent class analysis of embedded repeated measurements: An application to the National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society, Series A*, 165, 1.
- CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 397, 46-51.
- COHANY, S., POLIVKA, A. et ROTHGEB, J. (1994). Revisions Current Population Survey. Employment and Earnings BLS Report.
- COHEN, J.A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 37-46, 1960.
- DIPPO, C., POLIVKA, A., CREIGHTON, K., KOSTANICH, D. et ROTHGEB, J. (1994). Redesigning a Questionnaire for Computer-Assisted Data Collection: The Current Population Survey Experience.
- KOSTANICH, D., et CAHOON, L. (1994). CPS Bridge Team Technical Report 3: Effect of Design Differences Between the Parallel Survey and the New CPS. Rapport non-publié.
- MILLER, S. (1994). What Would the Unemployment Rate Have Been Had the Redesigned Current Population Survey Been in Place from September 1992 to December 1993? A Measurement Error Analysis. CPS Bridge Team Technical Report 1.
- POLIVKA, A. (1994). Comparisons of Labor Force Estimates from the Parallel Survey and the CPS During 1993: Major Labor Force Estimates. CPS Overlap Analysis Team Technical Report 1.
- POTERBA, J., et SUMMERS, L. (1995). Unemployment benefits and labor market transitions: A multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.
- POULSEN, C.S. (1982). *Latent Structure Analysis with Choice Modeling Applications*. Doctoral dissertation, Wharton School, University of Pennsylvania.
- ROTHGEB, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality. U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, Avril 6.
- SINCLAIR, M., et GASTWIRTH, J. (1998). Estimations des erreurs de classification dans l'enquête sur la population active et analyse de leur incidence sur les taux de chômage publiés. *Techniques d'enquête*, 24, 171-183.
- THOMPSON, J. (1994). Mode Effects Analysis of Labor Force Estimates. CPS Overlap Analysis Teach Rapport technique, 3.
- VAN DE POL, F., et DE LEEUW, J. (1986). A Latent Markov Model to Correct for Measurement Error. *Sociological Methods and Research*, 15, 1-2, 118-141.
- VERMUNT, J. (1997). *ÉEM: A General Program for the Analysis of Categorical Data*. Tilburg, University.
- WIGGINS, J.M. (1973). Panel Analysis. Latent Probability Models for Attitude and Behavior Processing. Elsevier S.P.C., Amsterdam.

Numéro de la question	Questions(s) d'origine provenant du questionnaire de la CPS	La réponse à la question composée est affirmative si la réponse à la question d'origine est....
Questionnaire révisé		

01	Q19 : Qu'avez-vous fait le plus la SEMAINE DERNIERE?	Q19 : Toute réponse, sauf « travailler »
02	Q20 : Avez-vous fait du travail la SEMAINE DERNIERE, sans compter le travail dans la maison?	Q20 : Non
	Q21 : La SEMAINE DERNIERE, aviez-vous un emploi ou une entreprise dont vous vous êtes absente(e) temporairement ou dont vous avez été mis(e) à pied?	Non
03	Q22 : Avez-vous cherché du travail au cours des quatre dernières semaines?	Q22 : Oui ou la réponse à Q19 était LK (à la recherche d'emploi)
	Q22A : Qu'avez-vous fait pour trouver du travail au cours des quatre dernières semaines?	Q22A : Autre réponse que « rien »
04	Q22B : Avez-vous été capable de travailler la SEMAINE DERNIERE si on vous avait offert un emploi?	Oui ou Non, et la raison est « a déjà un emploi » ou « temporairement malade »
Questionnaire révisé		
N1	Q20 : La SEMAINE DERNIERE, avez-vous fait TOUT travail (soit) contre rémunération (soit en vue d'un bénéfice)?	Q20 : Non
N2	Q20B-a : La SEMAINE DERNIERE, (en plus de l'entreprise) aviez-vous un emploi, à temps plein ou à temps partiel? Veuillez inclure tout emploi dont vous êtes temporairement absent(e).	Q20B-a : Non
N3	Q22 : Avez-vous fait quoi que ce soit pour trouver du travail au cours des quatre dernières semaines?	Oui
N4	Q22A : Quelles sont toutes les choses que vous avez faites pour trouver du travail au cours des quatre dernières semaines?	Mention d'au moins une activité de recherche active
N5	Q22A-DK : Vous avez dit que vous avez essayé de trouver du travail. Qu'avez-vous fait pour chercher du travail? et Q22A-DK1 : Pouvez-vous m'en dire plus sur ce que vous avez fait pour chercher du travail?	Oui

Figure 5. Questions composées utilisées dans la variable recodée LOOKING pour les versions originale et révisée du questionnaire

Notat : Dans quelques cas, N2 était positive si la réponse à Q20B-a était « Invalide » ou « Incapable » et que la réponse à Q20A-1 : « Est-ce que votre invalidité vous empêche d'accepter n'importe quel type de travail pour les six prochains mois? » était « Non ».

5. CONCLUSION

Biemer et Bushery (2000) donnent certaines preuves que les taux d'exactitude de la classification des personnes en chômage étaient plus faibles pour la CPS remaniée de 1994 que pour le plan de sondage original utilisé avant 1994. Le présent article fournit des preuves supplémentaires de cette observation fondées sur une analyse de plus grande portée des données de la CPS couvrant la période allant de 1992 à 1994. Selon nos résultats, la probabilité de classer correctement les personnes en chômage a diminué de 5,6 points, pour passer de 79,1 % à 73,5 %. Nous estimons qu'environ 60 % de la réduction (3,4 points) sont dus à une augmentation de l'erreur de classification des personnes mises à pied, tandis que le reste (2,2 points) est dû à une augmentation de l'erreur de classification des personnes à la recherche d'un emploi.

Pour le questionnaire révisé, les classifications des personnes mises à pied (LAYOFF) et à la recherche d'emploi (LOOKING) sont toutes deux fondées sur cinq questions composées. Pour LAYOFF, deux questions initiales sur l'activité « La SEMAINE DERNIERE, avez-vous fait TOUT travail (soit) contre rémunération (soit en vue d'un bénéfice)? ». La contribution de cette composante à l'erreur de classification concernant LAYOFF est de l'ordre de 57 %, c'est-à-dire plus du double du taux correspondant pour cette question dans le questionnaire original. En outre, le taux d'erreur estimatif est important pour la question composée formée par les deux questions : « Votre employeur vous a-t-il donné une date de retour au travail? » et « Avez-vous des raisons de croire que vous serez rappele(e) au travail au cours des six prochains mois? ». Environ 34 % du taux d'erreur estimé pour la classification des personnes mises à pied sont imputables à cette combinaison. Puisqu'il n'existe aucune question correspondante dans le questionnaire original, la plupart de l'erreur de classification des personnes mises à pied au moyen du questionnaire révisé peut être associée à l'introduction de ces deux questions.

nouveau assez comparables. La composante N1 semble être une source importante d'erreur pour la classification LOOKING, comme cela était le cas pour la classification LAYOFF. Cependant, sa contribution est plus faible pour LOOKING que pour LAYOFF, soit 10 points comparativement à 25 points. Le contributeur le plus important à l'erreur pour la classification LOOKING semble être la question N3 qui est la cause de 64,5 % de l'erreur fondée sur l'analyse CCO et de 51,1 % de l'erreur fondée sur l'analyse des données de la CPS de 1994.

Donc, la question initiale sur la situation d'activité semble être problématique pour les deux versions du questionnaire. La MILCA donne à penser que les personnes à la recherche d'un emploi ainsi que les personnes mises à pied éprouvent certaines difficultés à répondre à la question « La SEMAINE DERNIÈRE, avez-vous fait TOUT travail (soit) contre rémunération (soit en vue d'un bénéfice)? ». Les modifications apportées à cette question en 1994 ne semblent avoir amélioré l'exactitude de la réponse pour ni l'une ni l'autre population.

Pour la catégorie LOOKING, il semble que la difficulté principale consiste à déterminer si les personnes réellement à la recherche d'un emploi ont fait des efforts de quatre quel type (recherche passive ou active) au cours des quatre dernières semaines en vue de trouver du travail. Selon les estimations du tableau 4, si un répondant est classé correctement comme ayant fait certains efforts, l'étape suivante du processus, c'est-à-dire déterminer si les efforts satisfont à la définition de la recherche active, ne pose pas de problème.

Selon l'approche utilisée pour LAYOFF, pour chaque année de référence, nous définissons une variable latente, W dans (8) et une variable indicatrice, R dans (9). Comme nous l'avons fait pour l'analyse de la classification LAYOFF, nous ajustons des modèles MILCA aux données et déterminons que le meilleur d'entre eux pour l'analyse est celui contenant la variable de groupement réponse avec ou sans procuration, P , et spécifiant des probabilités de changement de situation non stationnaires, des probabilités de réponse égales durant la période de référence, des probabilités de changement de situation hétérogènes dans les groupes et des probabilités de réponse hétérogènes. Ce modèle donne un ajustement adéquat aux données pour tous les mois visés par l'analyse (c'est-à-dire $p > 0,05$). Comme précédemment, nous incluons les résultats de l'Enquête parallèle aux fins de comparaison aux résultats de la CPS de 1994; toutefois, nous mettrons l'accent sur ces derniers.

Le tableau 4 donne les valeurs de p_i définies en (11) pour la classification LOOKING. Pour le questionnaire original, les principaux contributeurs à l'erreur de classification semblent être les questions O1 et O3, qui causent 31,5 % et 56,3 % de l'erreur de classification totale, respectivement. La question O2, qui était relativement problématique pour la classification LAYOFF, semble l'être moins pour la classification LOOKING. Alors qu'elle était la cause de 64,2 % de l'erreur estimée pour LAYOFF (ou 24,8 points du taux d'erreur), O2 n'est la cause que de 11,3 % de l'erreur estimée pour LOOKING (ou 2,5 points du taux d'erreur).

Pour le questionnaire révisé, les résultats de l'analyse des données de l'Enquête parallèle et de la CPS de 1994 sont de

Question	CPS de 1993	Enquête parallèle	CPS de 1994
	(version originale)	(version révisée)	(version révisée)
Questionnaire original	Taux d'erreur	Taux d'erreur	Taux d'erreur
O1	6,93	31,51	—
O2	2,49	11,34	—
O3	12,39	56,33	—
O4	0,18	0,83	—
Questionnaire révisé	—	—	—
N1	—	8,38	33,00
N2	—	0,00	0,00
N3	—	16,38	64,50
N4	—	0,46	1,81
N5	—	0,18	0,71
Total	22,00	25,39	25,39
			Pourcentage du total
			100,00

Contribution en pourcentage à l'erreur de classification de la catégorie LOOKING selon la question composée pour la CPS de 1993, l'Enquête parallèle et la CPS de 1994

Tableau 4

de 90 % de l'erreur de la classification LAYOFF est due à deux composantes, à savoir N1 et N4.

Nous avons répété l'analyse pour le questionnaire révisé sur les données de l'Enquête parallèle et avons obtenu des résultats fort semblables, les deux mêmes composantes étant à l'origine de plus de 90 % de l'erreur. Comme nous l'avons mentionné à la section 2, l'utilité de l'Enquête parallèle de 1993 comme indicateur de la qualité des données recueillies au moyen du questionnaire révisé est douteuse. Néanmoins, la concordance entre les résultats de cette enquête et ceux de la CPS de 1994 renforce les observations fondées sur l'analyse des données de la CPS de 1994.

Donc, la réduction de l'exactitude de la classification dans la catégorie LAYOFF observée pour le questionnaire révisé semble être due principalement aux erreurs de réponse à deux questions composées, à savoir N1, la question globale révisée « La SEMAINE DERNIÈRE, avez-vous fait TOUT travail (soit contre rémunération) soit en vue d'un bénéfice? » et N4, qui détermine si une personne déclarant une forme de mise à pied possède une date de retour au travail ou a des raisons de croire qu'elle sera rappelée au travail à une date ultérieure. D'après les estimations MLCA, presque 60 % de l'erreur de la classification LAYOFF révisée pourrait être due à N1, et environ 34 %, à N4.

4.2.2 Décomposition de la variable recodée LOOKING

Nous avons également appliqué le processus d'estimation décrit pour la variable LAYOFF à la variable recodée LOOKING (recherche d'un emploi). Notons que les questions composées O1, O2, N1 et N2 définies à la figure 5 pour LOOKING sont les mêmes que celles définies à la figure 4 pour LAYOFF. Puisque O1, O2 et N1 semblent être problématiques pour la classification LAYOFF, nous pourrions nous attendre à ce qu'elles le soient également pour LOOKING.

Tableau 3
Contribution en pourcentage à l'erreur de classification dans la catégorie LAYOFF selon la question composée pour la CPS de 1993, l'Enquête parallèle et la CPS de 1994

Question	CPS de 1993 (version originale)	Enquête parallèle (version révisée)	CPS de 1994 (version révisée)
Taux d'erreur	10,53	25,34	44,37
Pourcentage du total	27,20	52,26	100,00
Taux d'erreur	64,19	—	—
Pourcentage du total	1,74	—	—
O1	2,84	—	—
O2	2,35	—	—
O3	0,67	—	—
O4	—	—	—
Questionnaire révisé	—	—	—
N1	—	23,19	44,37
N2	—	0,00	0,00
N3	—	2,76	18,42
N4	—	41,52	41,52
N5	—	0,00	0,89
Total	38,39	100,00	100,00

référence, comportant chacune $K + 1 = 5$ classes latentes. Pour le questionnaire révisé, nous avons utilisé un modèle identique, sauf que chaque variable latente comptait $K + 1 = 6$ classes latentes.

Comme auparavant, le meilleur modèle MLCA pour cette analyse comprenait la variable de groupement réponse par procuration-sans procuration, P , et des probabilités de changement de situation non stationnaires spécifiques, des probabilités de réponse égales durant la période de référence, des probabilités de changement de situation hétérogènes dans les groupes et des probabilités de réponse hétérogènes. Le modèle donne un ajustement adéquat aux données pour tous les mois visés par l'analyse (c'est-à-dire $p > 0,05$).

Le tableau 3 résume les résultats de l'analyse. Dans la colonne intitulée « Pourcentage du total », nous présentons $p_k \times 100$ %, où

$$p_k = \frac{\pi_{k|k}}{\sum_{k=1}^K \pi_{k|k}} \quad (12)$$

est la proportion de l'erreur de classification due à la question composée k de la figure 4 et où les $\pi_{k|k}^{RLV}$ sont les estimations MLCA de $\pi_{k|k}^{RLV}$.

La contribution à l'erreur totale présentée au tableau 3 (colonne Pourcentage du total) est donnée par $p_k \times \Pr(A \neq 2 | X = 2)$, où p_k est donné par (12) et $\Pr(A \neq 2 | X = 2)$ est estimé d'après le tableau 2, comme étant égal à 1 moins le taux d'exactitude pour la catégorie LAYOFF. Pour le questionnaire original, les composantes qui contribuent le plus à l'erreur de classification pour la catégorie LAYOFF sont les questions O2 (64,2 %) et O1 (27,2 %). Ensemble, ces deux questions expliquent plus de 90 % de l'erreur de la classification LAYOFF.

Pour le questionnaire révisé, les estimations calculées d'après les données de la CPS de 1994 indiquent que plus

Numéro de la question composée	[TRADUCTION] (Question(s) d'origine provenant du questionnaire de la CPS)	La réponse à la question composée est affirmative si la réponse à la question d'origine est...

Questionnaire original		
O1	Q19 : Qu'avez-vous fait le plus la SEMAINE DERNIÈRE?	Q19 : Toute réponse, sauf « travailler »
O2	Q21 : La SEMAINE DERNIÈRE, aviez-vous un emploi ou une entreprise travail dans la maison?	Oui
O3	Q21A : Pourquoi vous êtes-vous absenté(e) de votre travail la SEMAINE DERNIÈRE?	Mise à pied temporaire (moins de 30 jours) ou mise à pied indéfinie (30 jours ou plus, ou pas de date précise de rappel)
O4	Q22B : Avez-vous été capable de travailler la SEMAINE DERNIÈRE si on vous avait offert un emploi?	Oui
Questionnaire révisé		
N1	Q20 : La SEMAINE DERNIÈRE, avez-vous fait TOUT travail (soit) contre rémunération (soit en vue d'un bénéfice)?	Non
N2	Q20B-a : La SEMAINE DERNIÈRE, (en plus de l'entreprise) aviez-vous un emploi, à temps plein ou à temps partiel? Veuillez inclure tout emploi dont vous êtes temporairement absenté(e).	Toute réponse, sauf « retraité(e) », « handicapé(e) » ou « incapable de travailler »
N3	Q20B-b : La SEMAINE DERNIÈRE, étiez-vous en mise à pied temporaire?	Q20B-b : Oui
N3	Q20B-1 : Quelle est la raison principale pour laquelle vous êtes absenté(e) du travail la SEMAINE DERNIÈRE?	Q20B-1 : « En mise à pied » ou « conjoncture économique défavorable »
N4	Q21 : Votre employeur vous a-t-il donné une date de retour au travail?	Q21 : Oui
N4	Q21A : Avez-vous des raisons de croire que vous serez rappelé(e) au travail d'ici six mois?	Q21A : Non et ou Q21A : Oui
N5	Q21A-1 : Avez-vous pu retourner au travail la SEMAINE DERNIÈRE si vous aviez été rappelé(e)?	Q21A-1 : Oui ou Q21A-1 : Non et ou Q21A-2 : Quelle en est la raison?

Figure 4. Questions composées utilisées dans la variable recodée LAYOFF pour les versions originale et révisée du questionnaire

Le k^e terme de la somme peut être interprété comme étant la contribution de la question \bar{Q}_k à la probabilité d'être classé incorrectement, étant donné une mise à pied réelle. Pour estimer les composantes de (9) par la MICA, nous définissons une variable de classification, R , par analogie à W pour les valeurs observées de \bar{Q}_k ; c'est-à-dire

$$R = \begin{cases} 0 & \text{si } \bar{Q}_1 = 2, \bar{Q}_2 = 2, \dots, \bar{Q}_K = 2 \\ 1 & \text{si } \bar{Q}_1 = 1, \bar{Q}_2 = 2, \dots, \bar{Q}_K = 2 \\ \dots \text{etc.} \dots & \\ K & \text{si } \bar{Q}_1 = 1, \bar{Q}_2 = 1, \dots, \bar{Q}_K = 1. \end{cases} \quad (10)$$

Soit $\pi_{k|K}^{R|W}$, $\Pr(R=k|W=K)$. Alors, pour $k > 0$, nous pouvons écrire

$$\pi_{k|K}^{R|W} = \Pr(\bar{Q}_1 = 1, \dots, \bar{Q}_{k-1} = 1, \bar{Q}_k = 2 | W = K). \quad (11)$$

Donc, nous pouvons déterminer la contribution à l'erreur de chaque question concernant la mise à pied (LAYOFF) d'après les probabilités données par (11). Pour estimer les probabilités $\pi_{k|K}^{R|W}$, nous ajustons des modèles MICA pour les mêmes données provenant de la CPS de 1993 et de 1994 que celles utilisées lors de l'analyse précédente, et nous répétons l'analyse sur les données de l'enquête parallèle de 1993. Les données provenant de la CPS de 1992 et de 1995 n'ont pas été incluses dans la présente analyse. Les modèles MICA utilisés sont comparables à ceux décrits dans l'analyse correspondant aux tableaux 1 et 2. Autrement dit, nous avons utilisé des données recueillies pendant trois mois consécutifs et estimé les composantes de (10) pour 10 intervalles consécutifs, les chevauchants, pour chaque année (c'est-à-dire de janvier à mars, de février à avril, et ainsi de suite jusqu'en octobre à décembre). Pour le questionnaire original, nous avons spécifié dans le modèle trois variables latentes continues correspondant aux trois mois compris dans une période de

individuelles utilisées pour classer une personne dans la catégorie LAYOFF. La méthode employée pour cela est comparable à l'approche MLCA suivie antérieurement pour estimer l'erreur de classification agrégée. Nous décrirons cette technique en ce qui a trait à la classification des personnes mises à pied (LAYOFF), mais nous l'appli- quons subseqüemment à la décomposition de l'erreur dans les processus de classification des mises à pied (LAYOFF).

Premièrement, nous combinons les questions de la figure 3 en utilisant des opérateurs logiques, tels que « et », « ou » et « si-alors-autrement », etc. pour constituer un ensemble de questions « composées » dichotomiques ayant pour propriété que chaque question composée doit donner lieu à une réponse affirmative pour qu'une personne soit classée dans la catégorie LAYOFF en vertu du processus de classification de la CPS. Soit $\bar{Q}_k, k = 1, \dots, K$ les résultats des K questions composées qui ont été formées pour la classification LAYOFF, où $\bar{Q}_k = 1$ représente un résultat positif et $\bar{Q}_k = 2$, un résultat négatif. Alors, une personne participant à la CPS sera classée dans la catégorie LAYOFF si, et uniquement si, $\bar{Q}_k = 1$ pour $k = 1, \dots, K$. À la figure 4, nous définissons un ensemble de quatre questions composées pour le questionnaire révisé, étiquetées N1 à N5.

Pour chaque classification, \bar{Q}_k , il existe une classification réelle, inobservable (latente) correspondante, T_k , définie par analogie à \bar{Q}_k ; autrement dit, un individu est réellement dans la catégorie des personnes mises à pied selon la définition de la CPS si, et uniquement si, $T_k = 1, k = 1, \dots, K$. Puis, nous utilisons la méthode MLCA pour estimer les taux d'erreur de classification pour chaque question composée \bar{Q}_k en traitant ces questions comme des

indicateurs des caractéristiques latentes réelles inconnues, T_k . La probabilité d'une erreur concernant la classification LAYOFF peut s'écrire

$$\Pr(\bar{Q}_k = 2 \text{ pour un certain } k, k = 1, \dots, K | T_k = 1, k = 1, \dots, K) \quad (7)$$
$$W = \begin{cases} 0 & \text{si } T_1 = 2, T_2 = 2, \dots, T_K = 2 \\ 1 & \text{si } T_1 = 1, T_2 = 2, \dots, T_K = 2 \\ \dots \text{etc.} & \\ K & \text{si } T_1 = 1, T_2 = 1, \dots, T_K = 1. \end{cases} \quad (8)$$

Par exemple, $W = 0$ si le profil de réponse réel d'une personne aux questions O1 à O4 est (2,2,2,2), $W = 1$ si le profil réel de réponse est (1,2,2,2) et ainsi de suite. Notons que $W = K$ correspond à une mise à pied réelle. Donc, pour le questionnaire original, $W = 0, \dots, 4$ et pour le questionnaire révisé, $W = 0, \dots, 5$.

Pour décomposer la probabilité donnée par (7) en composantes individuelles pour la question composée, \bar{Q}_k , nous récrivons (7) en fonction des probabilités d'erreur associées à chaque question composée. Donc, nous pouvons montrer que (7) peut se réécrire sous la forme

$$\Pr(\bar{Q}_1 = 1, \dots, \bar{Q}_{k-1} = 1, \bar{Q}_k = 2 | W = K). \quad (9)$$

Questionnaire original	Énoncé de la question [TRANSCRIPTION]
Q19	Qu'avez-vous fait le plus la SEMAINE DERNIÈRE?
Q20	Avez-vous fait du travail la SEMAINE DERNIÈRE, sans compter le travail dans la maison?
Q21	La SEMAINE DERNIÈRE, avez-vous un emploi ou une entreprise dont vous êtes absent(e) temporairement ou dont vous avez été mis(e) à pied?
Q21A	Pourquoi vous êtes-vous absent(e) de votre travail la SEMAINE DERNIÈRE?
Q22E	Auriez-vous été capable de travailler la SEMAINE DERNIÈRE si on vous avait offert un emploi?
Questionnaire révisé	
Q20	La SEMAINE DERNIÈRE, avez-vous fait TOUT travail (soit contre rémunération (soit en vue d'un bénéfice)?
Q20B-a	La SEMAINE DERNIÈRE, (en plus de l'entreprise) avez-vous un emploi, à temps plein ou à temps partiel? Veuillez inclure tout emploi dont vous étiez temporairement absent(e).
Q20B-b	La SEMAINE DERNIÈRE, étiez-vous en mise à pied?
Q20B-1	Quelle est la raison principale pour laquelle vous vous êtes absent(e) du travail la SEMAINE DERNIÈRE?
Q21	Voire employeur vous a-t-il donné une date de retour au travail?
Q21A	Avez-vous des raisons de croire que vous serez rappelé(e) au travail d'ici six mois?
Q21A-1	Auriez-vous pu retourner au travail la SEMAINE DERNIÈRE si vous avez été rappelé(e)?
Q21A-2	Quelle en est la raison?

Figure 3. Composantes principales de UEM (en chômage) pour les questionnaires original et révisé

Tableau 1
Comparaison des probabilités de réponse concernant l'activité pour les questionnaires original et révisé de la CPS

Catégorie	Catégorie	Original	Révisé	Original – Révisé	E.T.
EMP	EMP	98,68	98,84	-0,15	0,40
	UEM	0,42	0,39	0,03	0,40
	NLF	0,90	0,78	0,13	0,16
UEM	EMP	8,23	10,57	-2,34*	0,45
	UEM	79,06	73,50	5,56*	0,54
	NLF	12,71	15,93	-3,32*	0,26
NLF	EMP	2,14	1,99	0,15	0,36
	UEM	1,43	1,56	-0,13	0,33
	NLF	96,43	96,45	-0,02	0,18

* Significatif à $\alpha = 0,001$.

Tableau 2
Comparaison de deux sous-catégories de chômeurs pour les questionnaires original et révisé

Catégorie réelle	Catégorie	Original	Révisé	Original – Révisé	E.T.
UEM – LAYOFF	EMP	16,32	26,67	-10,35*	0,91
	UEM – Layoff	61,30	55,63	5,66*	1,03
	NLF	17,61	8,41	9,20*	0,45
UEM – LOOKING	EMP	7,03	7,51	-0,48	0,29
	UEM – Layoff	1,03	0,65	0,38	0,26
	NLF	13,94	74,61	3,39*	0,21

* Significatif à $\alpha = 0,001$.

Pour les personnes mises à pied, l'introduction du questionnaire révisé semble avoir fait baisser l'exactitude de la classification de 5,66 points, en moyenne, pour passer de 61,30 % à 55,63 %. Cependant, les profils d'erreur de classification ont également changé. Pour le questionnaire original, la probabilité qu'une personne mise à pied soit classée incorrectement dans la catégorie des personnes à la recherche d'un emploi est estimée à environ 18 %. L'estimation correspondant pour le questionnaire révisé est plus de deux fois faible, soit 8,5 %. En outre, les données laissent entendre que la classification incorrecte des chômeurs mis à pied dans la catégorie des personnes occupées ou des personnes inactives a augmenté de 10,35 points et 4,52 points, respectivement.

Considérons maintenant les personnes qui sont réellement à la recherche d'un emploi dans la partie inférieure du tableau 2. D'après le modèle MLCA, l'exactitude de la classification après le remaniement de la CPS a diminué significativement, pour passer de 78,00 % à 74,61 %. La plupart des erreurs de classification sont dues à la classification de personnes à la recherche de travail dans la catégorie des personnes inactives, situation qui se produirait, par exemple, si les questions concernant les activités de recherche actives et passives d'un emploi sont susceptibles

de produire des réponses erronées. Pour examiner plus en détail ce résultat, nous avons analysé chacune des questions utilisées pour déterminer la variable recodée LOOKING qui est décrite à la section suivante. Pour commencer, nous considérons les sources de l'erreur de classification des mises à pied (LAYOFF), puis nous étudions les sources de l'erreur de classification des personnes à la recherche de travail (LOOKING).

4.2 Questions particulières causant la réduction de l'exactitude de la classification des mises à pied (LAYOFF)

4.2.1 Décomposition de la variable recodée LAYOFF

Les personnes qui participent à la CPS sont classées dans la catégorie des mises à pied (LAYOFF) d'après leurs réponses à cinq questions du questionnaire original et huit questions du questionnaire révisé. Ces questions sont énumérées à la Figure 3. Nous commençons par déterminer quelles questions ou combinaisons de questions contribuent le plus au taux d'erreur observé au tableau 2 pour la variable recodée LAYOFF, puis nous montrons comment les modèles MLCA peuvent être appliqués pour estimer la contribution à l'erreur de classification des questions

Leurs résultats empiriques montrent que cette variable est fortement corrélée non seulement à l'exactitude de la déclaration, mais aussi à la situation courante d'activité et aux changements de situation d'activité d'un mois à l'autre. Par exemple, les réponses étaient nettement moins exactes pour le groupe PROXY que pour le groupe SELF et, de surcroît, le taux de chômage était un peu plus élevé pour le premier que pour le second.

Le modèle MLCA permet aussi la variation des probabilités de changement de situation en fonction de P (hétérogénéité au sein du groupe) et en fonction de la période de référence (non-stationnarité). En outre, le modèle suppose que les probabilités de réponse $\pi_{a|px}$, $\pi_{b|px}$, et $\pi_{c|px}$ sont hétérogènes par rapport au groupe, mais sont égales pour les trois mois compris dans l'intervalle de temps. Nous obtenons donc le modèle suivant pour décrire les probabilités des cellules du tableau PABC :

$$(6) \quad \pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_x^y \pi_{d|px}^z \pi_{a|px}^x \pi_{b|py}^y \pi_{c|pz}^z$$

où $\pi_{a|px}^x = P(A=b|P=p, X=x) = y$ avec des définitions semblables pour $\pi_{b|py}^y$ et $\pi_{c|pz}^z$. Autrement dit, les trois ensembles de probabilités de réponses sont égaux à $\pi_{a|px}$.

Notons qu'aux fins de la présente analyse, nous nous concentrons sur les probabilités globales de réponse associées aux questionnaires révisé et original et non à la variation des taux d'erreur selon le groupe de réponse par procuration. Par conséquent, notre analyse est axée sur l'exactitude globale de la réponse, c'est-à-dire $\pi_{a|x}^x$ ou la probabilité moyenne de réponse pour les quatre niveaux combinés de P .

4. COMPARAISON DES PROBABILITÉS D'ERREUR DE CLASSIFICATION DES QUESTIONNAIRES RÉVISÉ ET ORIGINAL

4.1 Réduction de l'exactitude de la classification des chômeurs (UEM) pour le questionnaire révisé

Comme nous l'avons mentionné à la section 2, les ensembles de données de la CPS analysés sont ceux correspondant aux échantillons mensuels pour la période allant d'août 1992 à mai 1995. La figure 2 montre comment cette période a été subdivisée en 30 intervalles de trois mois chevauchants : 15 pour le questionnaire original et 15 pour le questionnaire révisé. Les intervalles sont numérotés dans le tableau aux fins de future référence. Par exemple, l'intervalle 1 couvre la période d'août 1992 à octobre 1992 durant laquelle on a utilisé le questionnaire original. Par conséquent, il peut fournir une estimation des probabilités de réponse, $\pi_{a|x}^x$, pour le modèle (6). Puisqu'il existe 30 intervalles de temps sur la période complète de 34 mois

soit occupée durant la période 3. Notons que, si $\lambda_1 = \lambda_2 = 1$, l'hypothèse markovienne tient exactement et l'écart de λ_1 et λ_2 par rapport à 1 est d'autant plus grand que l'écart des données par rapport à l'hypothèse markovienne est important. Biemer et Bushery ont constaté que, sur une fourchette de valeurs assez grande pour λ_1 et λ_2 , le biais absolu des estimations MLCA de l'exactitude de la classification du chômage n'est jamais supérieur à 3 points. Par exemple, dans le cas extrême d'une violation de l'hypothèse markovienne, la valeur prévue d'une estimation MLCA de l'exactitude de la classification dans la catégorie chômage serait de 77 % quand la valeur réelle du paramètre est de 80 %. Leurs résultats donnent à penser que, pour l'application à la CPS, la MLCA est relativement robuste aux violations de l'hypothèse markovienne.

Bien qu'il soit quasiment impossible de prouver leur validité, les estimations MLCA de l'erreur peuvent être assez utiles pour repérer les questions de l'enquête susceptibles de produire des erreurs de classification, c'est-à-dire les questions imparfaites. Ainsi, Biemer (2004) et Biemer et Wiesen (2002) démontrent l'utilité de la méthode MLCA pour repérer les questions problématiques et les défauts du processus de classification dans le cas d'enquêtes à grande échelle. Bien qu'on ignore la mesure dans laquelle les hypothèses de la MLCA peuvent être violées, l'utilité de la méthode pour l'exploration de plusieurs problèmes importants de conception des questionnaires a été bien décrite. Dans la présente application, nous utilisons la MLCA pour élaborer et tester des hypothèses quant aux sources de l'anomalie décrite par Biemer et Bushery pour le remaniement de la CPS de 1994.

Le modèle MLCA utilise ici est essentiellement le même que celui choisi par Biemer et Bushery pour leur analyse. Pour tenir compte de l'hétérogénéité de la population, ils ont considéré plusieurs variables explicatives démographiques et autres susceptibles d'être fortement corrélées à l'erreur de classification. Parmi ces variables, celle qui a donné les meilleurs résultats est l'indicateur de réponse par procuration ou sans, représenté par P , où

- | | | | |
|---|---|---|-------|
| 1 | si les trois interviews sont menées sans procuration (SELF) | } | $P =$ |
| 2 | si deux des trois interviews sont menées sans procuration (MOSTLY SELF) | | |
| 3 | si deux des trois interviews sont menées par procuration (MOSTLY PROXY) | | |
| 4 | si les trois interviews sont menées par procuration (PROXY) | | |

recueillies durant les premiers trimestres de 1993, 1995 et 1996, respectivement. Ils ont également réalisé plusieurs types d'analyses portant sur les données de réinterview non rapprochées de la CPS pour la même période. L'analyse des données de réinterview fournit une autre approche d'estimation de l'erreur de classification de la CPS, ainsi que des preuves de la validité de l'approche MLCA. L'évaluation de la validité de la MLCA de Biemer et Bushery reposait sur cinq critères, à savoir 1) les diagnostics du modèle, 2) la qualité de l'ajustement du modèle d'une année de référence à l'autre de la CPS, 3) la concordance entre le modèle et les estimations test-retest des probabilités de réponse, 4) la concordance entre le modèle et les estimations test-retest de l'incohérence et 5) la plausibilité des profils de l'erreur de classification. La méthode MLCA a donné de bons résultats pour les cinq tests. Ainsi, un même modèle a donné le meilleur ajustement aux données pour chaque année de référence, la concordance entre les estimations de la fiabilité par la méthode des classes latentes et celles obtenues par la méthode habituelle de test-retest était bonne, et les taux estimatifs d'erreur concordaient avec ceux calculés lors d'études antérieures, notamment celles de Chua et Fuller (1987, d'Abowd et Zellner 1985, de Porterba et Summers 1995 et Shincalir et Caswirth 1998).

De toute évidence, il est fort peu probable que l'hypothèse markovienne tienne pour les données sur la population active. Par exemple, les personnes en chômage durant les mois 1 et 2 d'une période de trois mois consécutifs pourraient ne pas avoir la même probabilité d'être en chômage durant le troisième mois que celles devenues chômeuses durant le deuxième mois. Le premier groupe pourrait contenir un plus grand nombre de personnes chroniquement en chômage que celui entrant dans la catégorie des chômeurs le deuxième mois. De surcroît, ce dernier groupe pourrait contenir une plus forte proportion de personnes temporairement sans travail en raison d'un changement d'emploi. Biemer et Bushery ont considéré les conséquences des estimations MLCA de l'erreur de classification quand l'hypothèse markovienne est violée.

Au moyen d'une simulation, ils ont constaté que le biais dans les estimations MLCA des probabilités de classification dépend de la gravité des écarts des données de la CPS par rapport à l'hypothèse markovienne. Ils ont défini deux paramètres, λ_1 et λ_2 , qui sont les ratios des probabilités conditionnelles. λ_1 est le ratio de la probabilité qu'une personne ayant le profil (EMP, UEM) durant les périodes 1 et 2, respectivement, soit occupée durant la période 3, et λ_2 est le ratio de la probabilité qu'une personne ayant un profil (EMP, EMP) soit occupée durant la période 3. Pareillement, λ_2 est le ratio de la probabilité qu'une personne dont le profil est (UEM, UEM) soit occupée durant la période 3 à la probabilité qu'une personne dont le profil est (EMP, UEM) soit occupée durant la période 3.

avec des définitions analogues pour les indicateurs de réponses $B_{g,t}$ et $C_{g,t}$ pour les périodes 2 et 3, respectivement. Grâce à une extension de la notation établie plus haut, nous représentons les probabilités de réponse dans chacune de ces classifications par $\pi^{alg_x} = \Pr(A = a | X = x)$, avec des définitions analogues pour π^{blgy} et π^{clgz} . Donc, $\pi^{a=1|g,x=2}$ est la probabilité qu'en vertu de la CPS, une personne du groupe g soit classée comme étant occupée ($A = 1$) quand sa situation réelle est en chômage – mise à pied ($X = 2$). Pareillement, $\pi^{a=2|g,x=2}$ est la probabilité qu'en vertu de la CPS, une personne du groupe g soit classée comme étant en chômage – mise à pied.

Enfin, nous supposons que

$$(3) \qquad \pi^{a,b,c|g,x,y,z} = \pi^{alg_x} \pi^{blgy} \pi^{clgz}$$

ou que l'erreur de classification concernant la situation d'activité observée est indépendante de l'un des trois mois de référence à l'autre.

Les classifications de l'activité selon la CPS pour chaque mois d'un intervalle de trois mois consécutifs sont les variables de résultat de notre analyse. Soit A , B et C , les classifications observées et soit X , Y et Z , les classifications réelles (non observées) pour les mois 1, 2 et 3, respectivement. Soit G , une variable de groupement (ou de stratification) que nous définirons plus tard. Sous ces hypothèses, nous pouvons écrire la probabilité de classer un membre de l'échantillon de la CPS dans la cellule (g, a, b, c) du tableau $GABC$ comme suit :

$$(4) \qquad \pi^{gabc} = \sum_{x,y,z} \pi^g \pi^{x|g} \pi^{y|gx} \pi^{z|gy} \pi^{a|gx} \pi^{b|gy} \pi^{c|gz}.$$

Les extensions à plus d'une variable de groupement sont simples.

Sous échantillonnage multinomial, la fonction de vraisemblance pour le tableau $GABC$ est

$$(5) \qquad \Pr(GABC) = C \prod_{g,a,b,c} \pi^{gabc}$$

où C est la constante multinomiale et Π représente le produit des termes sur les indices inférieurs g, a, b et c . Sous les hypothèses émis plus haut, les paramètres du modèle peuvent être estimés par les méthodes d'estimation du maximum de vraisemblance. Van de Pol et de Leeuw (1986) donnent la formule pour l'application de l'algorithme E-M pour estimer les paramètres de ce modèle et décrivent les conditions de leur estimabilité. Nous avons utilisé le logiciel EBM (Vermunt 1997) pour ajuster les modèles MLCA.

Pour étudier la validité des estimations MLCA en vue d'analyser l'erreur de classification de l'activité selon la CPS, Biemer et Bushery ont analysé les données de la CPS

NLF n'était pas nécessaire pour notre analyse, si bien que nous avons regroupé les sous-catégories de la variable à sept catégories pour en faire une variable à quatre catégories correspondant à EMP, UEM-LOOKING, UEM-LAYOFF et NLF. La correspondance entre les variables à trois et à quatre catégories est présentée à la figure 1.

3. MODÈLES DE CLASSES LATENTES POUR L'ERREUR DE CLASSIFICATION DE LA CPS

Les modèles markoviens de classes latentes ont été proposés pour la première fois par Wiggins (1973) et par Pousins (1982). Van de Pol et de Leeuw (1986) ont établi les conditions sous lesquelles le modèle est identifiable et ont donné d'autres conditions d'estimabilité des paramètres du modèle. À la présente section, nous décrivons le modèle de base proposé par Biemer et Bushery (2000) et ses extensions en vue de l'application à l'analyse courante.

Subdivisons la population cible de la CPS en L groupes (par exemple, selon l'âge, la race ou le sexe). Van de Pol et de Leeuw (1986) ont établi les conditions sous lesquelles le modèle est identifiable et ont donné d'autres conditions d'estimabilité des paramètres du modèle. À la présente section, nous décrivons le modèle de base proposé par Biemer et Bushery (2000) et ses extensions en vue de l'application à l'analyse courante.

$$X_{g_i} = \begin{cases} 1 & \text{si la personne } (g, i) \text{ est occupée durant la période } i \\ 2 & \text{si la personne } (g, i) \text{ est en chômage - mise à pied durant la période } i \\ 3 & \text{si la personne } (g, i) \text{ est en chômage - à la recherche d'emploi durant la période } i \\ 4 & \text{si la personne } (g, i) \text{ est inactive durant la période } i \end{cases}$$

$$A_{g_i} = \begin{cases} 1 & \text{si la personne } (g, i) \text{ est classée dans la catégories EMP à la période } i \\ 2 & \text{si la personne } (g, i) \text{ est classée dans la catégories UEM - LAYOFF à la période } i \\ 3 & \text{si la personne } (g, i) \text{ est classée dans la catégories UEM - LOOKING à la période } i \\ 4 & \text{si la personne } (g, i) \text{ est classée dans la catégories NLF à la période } i \end{cases}$$

C'est-à-dire qu'à la période 3, la situation réelle d'un individu ne dépend pas de la situation durant la période 1, une fois que la situation à la période 2 est connue. Une autre interprétation est que la situation courante, sachant la situation durant la période précédente, ne dépend pas du changement de situation de la période précédente.

Considérons maintenant les classifications de l'activité d'après la CPS, représentées par A_{g_i} , B_{g_i} , et C_{g_i} pour les périodes 1, 2 et 3, respectivement, où

$$\pi^{z|gxy} = \pi^{z|gxy}$$

$$\pi^{x|g} \pi^{y|g} \pi^{z|gxy} = \pi^{x|g} \pi^{y|g} \pi^{z|gxy}$$

Soit $\pi^{x,y,z|g}$, $\Pr(X=x, Y=y, Z=z | G=g)$, $\pi^{y|gx}$, $\Pr(Y=y | X=x, G=g)$ et soit $\pi^{z|gxy}$, $\Pr(Z=z | Y=y, X=x, G=g)$. Alors, la probabilité qu'un individu du groupe g ait la situation d'activité x durant la période 1, y durant la période 2 et z durant la période 3 est $\pi^{x,y,z|g}$, qui peut s'écrire

avec des définitions analogues pour Y^{g_i} et Z^{g_i} pour les périodes 2 et 3, respectivement. Conformément aux conventions de la littérature sur l'analyse des classes latentes, nous laisserons tomber les indices inférieurs des variables pour simplifier la notation.

Figure 1. Association de la variable de situation d'activité recodée à sept catégories aux variables à trois et à quatre catégories utilisées pour l'analyse

Questionnaire original				Questionnaire révisé			
1. Travail — au travail	1. Occupé(e) — au travail	2. Occupé(e) — absent(e)	3. En chômage — mise à pied	1. EMP	2. UEM — LAYOFF	3. UEM — LOOKING	4. NLF
2. A un emploi — pas au travail	2. Occupé(e) — absent(e)	3. En chômage — mise à pied	4. En chômage — recherche d'emploi	1. EMP	2. UEM — LAYOFF	3. UEM — LOOKING	4. NLF
3. En chômage — mise à pied	3. En chômage — mise à pied	4. En chômage — recherche d'emploi	5. À la retraite — inactif(ve)	1. EMP	2. UEM — LAYOFF	3. UEM — LOOKING	4. NLF
4. En chômage — recherche d'emploi	4. En chômage — recherche d'emploi	5. À la retraite — inactif(ve)	6. Handicapé(e) — inactif(ve)	1. EMP	2. UEM — LAYOFF	3. UEM — LOOKING	4. NLF
5. Travail sans rémunération (moins de 15 heures dans une ferme ou une entreprise familiale) ou temporairement absent(e) d'un travail non rémunéré	6. Incapable d'accepter un emploi s'il en était offert un	7. Personne inactive	7. Autre — inactif(ve)	1. EMP	2. UEM — LAYOFF	3. UEM — LOOKING	4. NLF

dans la catégorie des chômeurs. Sous la définition du questionnaire révisé, une personne est en chômage uniquement si tous les points suivants sont vrais : 1) sans emploi, 2) activement à la recherche d'un emploi ou en mise à pied et s'attend à un rappel au travail dans les six mois à venir et 3) capable de travailler (sauf à cause d'une maladie temporaire éventuelle).

Mise à pied. Les personnes mises à pied sont définies comme étant des personnes qui n'exercent plus leur emploi et qui attendent un rappel au travail dans ce même emploi. Le questionnaire original ne tenait pas compte ou ne recueillait pas l'information sur l'attente d'un rappel. Cette lacune posait un problème, parce que, pour la plupart des gens, l'expression « mise à pied » pourrait signifier la cessation permanente d'un emploi, au lieu de la perte temporaire de travail que les économistes essaient de mesurer.

Méthodes de recherche d'emploi. Pour être comptée comme en chômage et à la recherche d'un emploi, une personne doit avoir entrepris une recherche active de travail au cours des quatre semaines qui ont précédé l'enquête. Le questionnaire révisé inclut une question un peu plus générale au sujet des méthodes de recherche d'emploi offrant des catégories de réponses élargies et structurées pour permettre aux intervieweurs d'enregistrer plus facilement les réponses en faisant la distinction entre les activités actives et passives de recherche d'emploi. En outre, il contient des questions supplémentaires de suivi pour les personnes qui répondent « rien » ou « ne sais pas ».

Semaine de référence. Alors que le questionnaire original faisait référence à la SEMAINE DERNIÈRE, la période de référence n'était jamais définie explicitement. Le questionnaire révisé fournit les dates précises de la semaine de référence. Nous reviendrons à ces changements plus tard, lorsque nous discuterons des différences d'erreur de classification et d'erreur de spécification entre les questionnaires révisé et original.

Comme nous l'avons mentionné plus haut, Biemer et Bushery se sont concentrés sur une variable de situation d'activité recodée à trois catégories, à savoir personne occupée (EMP), personne en chômage (UEM) et personne inactive (NLF). Aux fins de la présente analyse, nous avons utilisé une variable recodée étendue, également disponible dans les fichiers de données à grande diffusion de la CPS. Cette variable subdivise la catégorie UEM en deux sous-catégories correspondant aux personnes mises à pied (LAYOFF) et aux personnes à la recherche d'un emploi (LOOKING). La variable à sept catégories subdivise également les catégories EMP et NLF en sous-catégories; cependant, ce niveau de détail dans les catégories EMP et

À cause d'un problème dû aux variables d'identification nécessaires pour coupler les ménages pour les mois de juin 1995 à décembre 1995, il n'a pas été possible d'inclure ces mois dans l'analyse. En outre, puisque nos conclusions resteront les mêmes si nous incluons les données pour 1996 ou les années subséquentes de la CPS, nous limitons l'analyse aux 15 mois qui précèdent et aux 15 mois qui suivent l'adoption du questionnaire révisé. Donc, pour la plupart de l'analyse qui suit, nous donnons les estimations moyennes pour la période allant d'août 1992 à décembre 1993 pour le questionnaire original et pour la période allant de janvier 1994 à mai 1995 pour le questionnaire révisé (il convient de souligner que, puisque nos estimations sont fondées sur une moyenne mobile de trois mois consécutifs, les estimations de l'erreur de classification tiennent compte des variations saisonnières des taux d'emploi et des probabilités de changement de situation).

2.2 Concepts relatifs à la population active

Le questionnaire révisé de la CPS a été introduit en 1994 pour améliorer la qualité globale des renseignements sur le marché du travail grâce à une révision extensive de questions et à l'utilisation de techniques informatiques pour la collecte des données. Suit la description de quelques concepts pertinents pour l'analyse courante sur lesquels renamenter le questionnaire a eu des répercussions.

Personne occupée. Le module sur la population active du questionnaire original débutait par la question « Qu'avez-vous fait le plus la SEMAINE DERNIÈRE (travailler, s'occuper de la maison, aller à l'école ou quelque chose d'autre)? » Les intervieweurs pouvaient modifier la partie entre parenthèses de cette question d'après l'âge du répondant. Dans certains cas, le mot « travail » ou « travailler » ne faisait pas partie de la question. Par exemple, si le répondant avait l'air d'un étudiant, l'intervieweur pouvait laisser tomber le mot « travailler ». Dans le questionnaire révisé, cette question a été remplacée par deux autres : « Est-ce qu'un membre de ce ménage possède une entreprise ou une ferme? » et « LA SEMAINE DERNIÈRE, avez-vous fait TOUT travail (soit) contre rémunération (soit en vue d'un bénéfice)? » où les parties entre parenthèses de la question sont lues si la réponse à la première question est « oui ». En outre, d'autres questions ont été ajoutées pour préciser si les gains ou les bénéfices proviennent de l'entreprise familiale ou de la ferme. Donc, le concept d'emploi du questionnaire révisé semble être un peu plus général et mieux défini que le celui du questionnaire original.

Personne en chômage. La définition du chômage est légèrement modifiée dans le questionnaire révisé. Dans le questionnaire original, les personnes attendant la révision d'un emploi pour commencer à travailler étaient classées

Notre analyse est axée sur une variable de classification de la population active dérivée de plusieurs questions de la section sur l'emploi du questionnaire de la CPS. Cette variable est souvent qualifiée de variable « recodée » de population active, puisqu'elle est déterminée en établissant la correspondance entre un profil des réponses aux questions sur l'emploi de la CPS et certaines catégories de population active, dont personne occupée – au travail, personne occupée – pas au travail, personne en chômage – à la recherche de travail, et ainsi de suite. Biernier et Bushery ont utilisé une variable de classification de la situation d'activité à trois catégories : personne occupée (EMP), personne en chômage (UEM) et personne inactive (NLF). Aux fins de la présente analyse, nous utilisons une variable à quatre catégories qui subdivise la catégorie UEM en personne en chômage – mise à pied (UEM-LAYOFF) et personne en chômage – à la recherche de travail (UEM-LOOKING). Cette subdivision constitue une première étape en vue d'isoler la source de l'inexactitude apparente de la classification des chômeurs. Cependant, comme nous le montrerons plus loin, une décomposition plus poussée de ces catégories sera nécessaire pour arriver à la cause profonde de l'erreur.

2. DONNÉES ET CONCEPTS

2.1 Les ensembles de données étudiés

Sauf pour l'Enquête parallèle, les données de la CPS que nous avons analysées ont été téléchargées à partir du site www.nber.org. Ce site Web contient des microdonnées de la CPS pour chaque mois, de janvier 1976 à décembre 2004. Nous avons appliqué la méthode d'analyse markovienne des classes latentes supplémentaires ni de données provenant d'une autre source que la CPS.

Durant l'analyse préliminaire, nous avons étudié l'exactitude de la classification de la CPS pour une période de six années, à savoir de janvier 1992 à décembre 1997. Cette

analyse avait pour but de déterminer si l'anomalie notée pour la première fois par Biernier et Bushery (2000) est un phénomène transitoire concernant uniquement les mois directement consécutifs à l'introduction du nouveau questionnaire ou s'il a persisté pendant quelques années après son adoption. Si elle est temporaire ou transitoire, l'anomalie pourrait être associée à des problèmes survenus durant l'adoption progressive du nouveau plan d'enquête, comme des problèmes de formation des intervieweurs ou des questions liées au démarrage de la collecte des données. Cependant, la preuve d'un effet persistant pourrait être le signe de problèmes concernant le plan d'enquête, liés par exemple au questionnaire, aux procédures d'interview ou à l'algorithme de recodage.

En appliquant la MLCA aux données de chaque mois de 1992 à 1997, nous avons déterminé que, bien que la réduction de l'exactitude varie quelque peu de mois en mois, elle persiste effectivement pour tous les mois postérieurs à l'introduction du questionnaire révisé. Les résultats confirment la conjecture émise par Biernier et Bushery d'un effet systémique éventuellement lié aux nouvelles questions sur le chômage adoptées en janvier 1994.

Faute d'espace, nous présentons ici les résultats pour une période de référence un peu plus courte que celle visée par l'analyse préliminaire, à savoir les années 1992, 1993, 1994 et 1995. Cette période comprend deux années où la CPS a été réalisée avec le questionnaire original et deux années où elle l'a été avec le questionnaire révisé. En outre, nous présentons certains résultats d'une application de la MLCA aux données de l'Enquête parallèle de 1993 qui peuvent être comparés à ceux obtenus pour la CPS principale.

Les ensembles de données analysés sont assez grands. Chaque estimation de l'erreur de classification que nous obtenons est fondée sur l'ensemble des ménages qui ont été interviewés durant la CPS pour les trois mois consécutifs. Sur l'ensemble des quatre années visées par notre analyse, le nombre total de ménages qui ont répondu chaque mois durant l'intervalle de trois mois varie d'environ 37 000 à plus de 40 000. Pour l'Enquête parallèle de 1993, le nombre correspondant de ménages est d'environ 10 000. Les estimations produites sont pondérées comme il convient pour les probabilités de sélection et d'autres redressements à posteriori et, par conséquent, reflètent les probabilités de réponses sur lesquelles reposent les estimations publiées de la CPS. Pour construire les poids, nous avons pris les poids moyens sur trois mois consécutifs et nous les avons combinés pour former un enregistrement longitudinal pour l'analyse (nous avons également réalisé des analyses non pondérées dont les résultats ont été fort semblables à ceux de l'analyse pondérée. Ceci donne à penser que le choix des poids a peu d'effets sur les résultats de l'étude).

« population active » et, donc, à celle d'un plus grand nombre de personnes qui ne travaillent pas dans la catégorie des personnes en chômage plutôt que dans celle des personnes inactives (voir, par exemple, Polivka 1994 et Rothgeb 1994).

Au départ, on a estimé que l'augmentation du taux de chômage due au nouveau questionnaire était d'environ un demi-point. Cependant, une analyse plus approfondie des données de l'Enquête parallèle a fait mettre en doute cette estimation et, selon un rapport publié ultérieurement, l'augmentation serait inférieure à un dixième de point (Polivka et Miller 1994). Nous examinons plus loin les réserves émises dans les rapports subséquents quant à l'utilité des données de l'Enquête parallèle pour évaluer l'effet du remaniement et nous en tenons compte dans notre analyse de ces données.

Une analyse indépendante réalisée par Biemer et Bushery (2000) a révélé une anomalie dans les données sur la population active de la CPS révisée qui n'avait été décelée lors d'aucune étude antérieure du remaniement de la CPS. Suivant une méthode d'analyse markovienne de classes latentes (MLCA), Biemer et Bushery ont comparé l'exactitude de la classification de la population active sous les plans de sondage original et révisé en comparant les taux d'erreur estimés d'après les données de la CPS pour 1993, 1995 et 1996. Ils ont défini l'exactitude de la classification de la population active comme étant la probabilité qu'une personne appartenant réellement à l'une des catégories de la population active, disons la catégorie *a*, soit classée comme faisant partie de *a* par la CPS; c'est-à-dire Pr (classée dans *a* | vraiment dans *a*). Par exemple, l'exactitude qu'une classification pour le chômage est la probabilité qu'une personne qui est vraiment en chômage, d'après la définition de la CPS, soit correctement classée comme étant chômeuse d'après les règles de classification de la CPS.

Au tableau 2 de leur article, Biemer et Bushery indiquent que l'exactitude de la classification des chômeurs a diminué de 5,7 points, pour passer d'environ 81,8 % (erreur-type = 0,90) en 1993 à 76,1 (erreur-type = 1,2) en 1995 et à 74,4 % (erreur-type = 1,2) en 1996. Ces résultats donnent à penser que la CPS remaniée donne lieu à un taux de classement erroné des personnes vraiment en chômage plus élevé que l'ancienne CPS. Les auteurs ont d'abord considéré que ce résultat pourrait être un artefact de la méthode d'analyse markovienne des classes latentes. Comme nous le montrons plus loin, la MLCA ne nécessite aucune mesure réelle ou « étalon d'or » de l'emploi pour estimer l'erreur de classification. Au contraire, la méthode s'appuie sur un modèle décrivant les variations mensuelles réelles de la situation d'activité, ainsi que le processus de classification des individus dans les catégories de population active. Il se pourrait que les changements d'activité qui s'écartent de la

spécification du modèle soient considérés comme des erreurs de classification dans le processus d'estimation.

Pour vérifier la validité des résultats de la MLCA, les auteurs ont réalisé une série d'analyses selon des méthodes classiques d'estimation, une analyse de l'erreur selon le groupe de population, des comparaisons des erreurs estimées à d'autres estimations publiées, ainsi qu'une simulation pour évaluer l'effet de l'échec du modèle sur les résultats. Par exemple, il existe des preuves que la fiabilité test-retest du classement dans la catégorie des chômeurs a diminué après le remaniement. Avant celui-ci, l'indice d'incohérence (mesure de non-fiabilité utilisée habituellement par le Census Bureau qui est égale à $1 - \kappa$, où κ est le coefficient kappa de Cohen (Cohen 1960)) pour la catégorie des chômeurs était de 30 %, en moyenne, pour la période 1992 – 1993. Après le remaniement, la valeur de l'indice a augmenté pour atteindre presque 40 % pour la période 1995 – 1996. Ces analyses corroborent l'affirmation de Biemer et Bushery selon laquelle l'exactitude de la méthode de classification des personnes en chômage a diminué après le remaniement.

Dans leur discussion des résultats, les auteurs spéculent que la diminution de l'exactitude de la classification pourrait témoigner d'un problème dû aux questions révisées sur le chômage. Autrement dit, ces questions pourraient donner lieu à une plus grande erreur de classification, donc, à une classification moins exacte. Ils envisagent aussi la possibilité d'une évolution des caractéristiques de la population de chômeurs de 1993 à 1995 et à 1996. Puisque le taux de chômage a baissé de 1993 à 1996, il se pourrait que les personnes qui seraient classées le plus exactement en vertu du système de la CPS aient quitté les rangs des chômeurs, en ne faisant plus partie de cette catégorie que les personnes qui seraient classées moins exactement. On pourrait vérifier cette hypothèse en estimant les taux d'exactitude des deux méthodes pour la même période de référence. L'Enquête parallèle offre un moyen de réaliser ce genre d'analyse.

Dans le présent article, nous poursuivons l'étude par la MLCA de la diminution de l'exactitude de la classification des chômeurs observée par Biemer et Bushery. Nous utilisons pour l'analyse des modèles MLCA fort semblables à ceux utilisés par ces auteurs pour estimer l'exactitude de la classification pour les versions originale et révisée du questionnaire de la CPS. Cependant, nous étendons la période de référence afin d'y inclure les 15 mois qui ont précédé et suivi l'adoption du questionnaire révisé, ce qui nous donne une période de référence de 30 mois consécutifs. En outre, nous utilisons les données de l'Enquête parallèle recueillies pour la période de janvier 1993 à décembre 1993 en vue de comparer l'exactitude du taux d'emploi calculé d'après les questionnaires original et révisé pour la même période.

Une analyse de l'erreur de classification pour les questions sur l'emploi révisées de la Current Population Survey

PAUL P. BIEMER¹

RÉSUMÉ

La réduction de l'inexactitude de la classification révisée des personnes en chômage dans la Current Population Survey (CPS) a été décrite dans Biemer et Bushey (2000). Dans le présent article, nous donnons des preuves supplémentaires de cette anomalie et essayons de découvrir la source de l'erreur grâce à une analyse étendue des données de la CPS recueillies avant et après le remaniement. L'article présente une approche novatrice de décomposition de l'erreur dans le cas d'un processus de classification complexe, comme la classification de la situation d'activité de la CPS, par une analyse markovienne de classes latentes (MILCA). En vue de déterminer la cause de la perte apparente d'exactitude de la classification des chômeurs, nous recensons les composantes clés du questionnaire qui déterminent les classifications et nous estimons la contribution de chacune à l'erreur totale du processus de classification. Ces travaux serviront d'orientation aux études futures des causes profondes des erreurs lors de la collecte de données sur la situation d'activité dans le cadre de la CPS, éventuellement au moyen d'expériences cognitives en laboratoire et (ou) d'essais sur le terrain.

MOTS CLÉS : Remaniement d'une enquête; erreur de mesure; analyse de classes latentes; taux de chômage; erreur de spécification.

1. INTRODUCTION

La Current Population Survey (CPS) est une enquête mensuelle réalisée par le U.S. Bureau of the Census pour le compte du Bureau of Labor Statistics (BLS) auprès d'environ 60 000 ménages. L'objectif principal de l'enquête est de fournir des estimations de l'emploi, du chômage et d'autres caractéristiques de la population active américaine dans son ensemble. Publiées mensuellement par le BLS, les estimations de l'effectif, de la composition et des caractéristiques dynamiques de la population active constituent l'un des principaux indicateurs économiques nationaux.

En janvier 1994, le questionnaire de la CPS a été révisé afin de tenir compte des recommandations faites par la Commission Levitan à la fin des années 1970, à savoir passer du mode d'interview avec questionnaire papier-crayon à des méthodes d'interview assistée par ordinateur et rendre plus claires certaines questions sur l'emploi, ainsi que pour plusieurs autres raisons décrites par Rothgeb (1994). L'objectif général du remaniement était d'améliorer la qualité des données recueillies dans le cadre de la CPS. Le questionnaire de l'enquête n'avait subi pour ainsi dire aucune modification depuis la dernière révision importante de 1967.

Le questionnaire révisé de la CPS a été adopté après d'importants travaux de recherche et de mise à l'essai qui ont débuté au milieu des années 1980. Le but des essais était d'évaluer la qualité et la faisabilité opérationnelle des diverses options de remaniement, y compris le passage d'un questionnaire papier-crayon à l'interview assistée par

ordinateur. Durant ces années de mise à l'essai, plus de 100 000 personnes ont été interviewées dans le cadre des diverses études réalisées (Rothgeb 1994). Le point culminant du programme d'étude du remaniement de la CPS a été une grande étude de portée nationale (appelée dans la littérature CATI/API Overlap ou CCO Field Test) qui a été réalisée en 1993. La composante principale de cet essai était une enquête assistée par ordinateur auprès d'environ 12 000 ménages avec application des procédures d'interview révisées de la CPS et du questionnaire révisé. Cette enquête, que nous appelons dans le présent article Enquête parallèle, a été réalisée de juillet 1992 à décembre 1993 concomitamment à la CPS ordinaire basée sur le questionnaire original. Ce plan à panel fractionné permet d'estimer l'effet du remaniement sur les estimations de la population active fondées sur la CPS.

Plusieurs articles et rapports publiés décrivent les résultats du CCO Field Test (Cohany, Polivka et Rothgeb 1994; Rothgeb 1994; Polivka 1994; Kostonich et Cahoon 1994; Miller 1994; Thompson 1994; Dippo, Polivka, Creighton, Kostonich and Rothgeb 1994). L'un des résultats importants de cette étude est que les taux de chômage et de participation au marché du travail fondés sur l'Enquête parallèle étaient plus élevés que ceux calculés d'après la CPS. Certaines modifications de la définition de l'emploi sont les causes principales des taux plus élevés associés au questionnaire révisé. Ce dernier repose sur une approche plus générale du travail et des activités de recherche d'emploi qui a tendance à aboutir à la classification d'un plus grand nombre de personnes dans la catégorie

Zheng et Little proposent une méthode fondée sur un modèle non paramétrique pour remplacer l'estimation d'Horvitz-Thompson d'un total dans le cas de l'échantillonnage à deux degrés avec échantillonnage PPT au premier degré. Il s'agit d'une prolongation de leurs travaux antérieurs constant à modéliser une variable de résultat y_i sous forme d'une fonction lisse de la probabilité de sélection π_i . Ils montrent comment ajuster le modèle et estimer le total au moyen d'une spline pénalisée, et élaborent aussi des méthodes de rechange d'estimation de la variance. Ils se servent de simulation pour comparer la méthode proposée à l'estimateur d'Horvitz-Thompson et à un estimateur assisté par modèle.

Liang et Kuk considèrent une autre méthode que l'approche classique d'estimation par la régression en population finie. Au lieu du modèle linéaire habituel, ils utilisent une fonction lisse arbitraire pour permettre une régression non linéaire, puis ils appliquent la technique des réseaux neuronaux bayésiens pour résoudre le problème. L'avantage de l'approche des réseaux neuronaux tient à ce qu'elle permet d'éviter le problème de la spécification erronée du modèle. Liang et Kuk appliquent une loi a priori à chaque connexion du réseau plutôt qu'uniquement au nombre d'unités cachées, comme cela est généralement le cas. Ceci leur permet de traiter uniformément la sélection de la structure du réseau et celle des variables auxiliaires. Enfin, ils traitent les valeurs aberrantes en introduisant une distribution à queue lourde pour modéliser les perturbations des données.

Dans le dernier article de ce numéro, Reiter recourt à l'imputation multiple pour résoudre simultanément les questions des données manquantes et du contrôle de la divulgation. L'idée fondamentale est de commencer par remplacer les données manquantes pour générer m ensembles de données complètes, puis de remplacer les valeurs délicates ou identifiantes comprises dans chaque ensemble de données complètes au moyen de r valeurs imputées. Ensuite, l'auteur élabore de nouvelles règles de combinaison pour obtenir des inférences valides d'après de tels ensembles de données multi-imputés. Ces règles tiennent compte des deux sources de variabilité dans les estimateurs ponctuels.

Finalement, le comité éditorial a tenu une réunion l'été dernier au Joint Statistical Meetings à Toronto. Il y a été suggéré d'avoir une section sur de brèves communications dans la revue. Celles-ci seraient de courts articles qui auraient environ quatre pages de format Techniques d'enquête. Les sujets possibles de communication pourraient inclure la présentation de nouvelles idées sans le développement complet d'un article régulier de brefs rapports sur des travaux empiriques et des discussions ou compléments à d'autres articles publiés dans la revue. Toutes les communications brèves seraient arbitraires, quoique le processus de revue pourrait être simplifié. J'espère que le nouveau format attirera plusieurs auteurs, et j'ai bien hâte de recevoir vos propositions de communications.

M.P. Singh

Dans ce numéro

Le présent numéro débute par un article de Paul Biemer, suivi d'une discussion. L'article donne des preuves d'une réduction de l'exacitude de la Current Population Survey (CPS) causée par le remaniement des questions sur l'emploi. Il s'agit d'une extension de l'étude réalisée antérieurement par Biemer et Bushery (2000). Dans l'article courant, l'auteur s'efforce de décrire la source de l'erreur grâce à une analyse approfondie des données de la CPS avant et après le remaniement. Il présente une nouvelle approche, fondée sur l'analyse markovienne de classes latentes. Ce travail a pour objectif d'orienter l'étude future des causes profondes des erreurs qui entachent la collecte des données sur la population active dans le cadre de la CPS. Les discussions concernant l'article ont été rédigées par Jerroen Vermunt, Stephen Miller et Anne Polivka, ainsi que Clyde Tucker.

Dans leur article, Gunning et Horgan proposent un nouvel algorithme pour la construction de limites de strate dans les populations asymétriques. Cet algorithme comprend l'utilisation d'une variable auxiliaire et produit des coefficients de variation de même valeur dans chaque strate pour cette variable auxiliaire. La méthode est fondée sur l'hypothèse selon laquelle la variable auxiliaire suit une loi de distribution uniforme. L'un de ses avantages tient au fait qu'elle est facile à appliquer en pratique. Au moyen d'une étude empirique, les auteurs montrent que l'algorithme proposé donne des résultats qui se comparent favorablement à ceux de la méthode de la fonction cumulative de la racine carrée des fréquences de Dalenius et Hodges (1957) et à ceux de l'algorithme de Lavallée et Hidiroglou (1988).

Hedlin et Wang considèrent le problème du biais causé par l'introduction d'information en retour provenant des enquêtes dans les bases de sondage. Ils examinent le biais produit par la mise à jour des données sur les décès dans une base de sondage utilisée pour la réalisation de cycles futurs de la même enquête. Ils quantifient ce biais et proposent un estimateur sans biais applicable dans cette situation. Les résultats théoriques présentés dans l'article sont illustrés grâce à une étude en simulation.

Dans leur article, Mudyk et Xie présentent les aspects assurance de la qualité (AQ) et contrôle de la qualité (CQ) de l'opération de reconnaissance intelligente de caractères du Recensement de l'agriculture du Canada de 2001. Ils montrent comment un bon plan d'AQ et de CQ a été élaboré pour assurer que les opérations de saisie du recensement produisent des données de la plus haute qualité possible. Les résultats d'une analyse de la qualité moyenne des données après contrôle soulignent l'importance qu'il y a à établir un plan d'AQ/CQ.

Park et Lee étudient les effets du plan pour les estimateurs pondérés de la moyenne et du total dans le cas d'enquêtes complexes. Plus précisément, ils décomposent l'effet du plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage à deux degrés. Partant de cette décomposition, ils illustrent plusieurs idées fausses concernant les effets du plan pour les estimateurs pondérés de la moyenne et du total à l'aide de divers exemples fondés sur des plans de sondage utilisés couramment.

Dans leur article, Beaumont et Alavi étudient un estimateur par la régression généralisée robuste. Ils recherchent d'autres options que le meilleur estimateur linéaire sans biais (BLU) optimal qui soient robustes à l'hypothèse que le plan de sondage est ignorable et (ou) aux erreurs de spécification du modèle. Dans la situation où l'hypothèse que le plan de sondage est ignorable pourrait ne pas tenir, ils proposent un estimateur par les moindres carrés obtenu par réajustement des poids de sondage vers la moyenne. Pour résoudre la question des erreurs de spécification du modèle, ils proposent un estimateur M généralisé pondéré pour réduire l'influence des unités dont le résidu pondéré de la population est important. Ils illustrent leurs résultats théoriques au moyen d'une étude en simulation.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 30, numéro 2, décembre 2004

TABLE DES MATIÈRES

Dans ce numéro	139
Article de discussion	
PAUL P. BIEMER	
Une analyse de l'erreur de classification pour les questions sur l'emploi révisées de la Current Population Survey	141
Commentaires:	
JEROEN K. VERMUNT	156
STEPHEN M. MILLER et ANNE E. POLIVKA	160
CLYDE TUCKER	167
Réponse de l'auteur	171
Articles réguliers	
PATRICIA GUNNING et JANE M. HORGAN	
Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques	177
DAN HEDLIN et SUOJIN WANG	
Introduction dans la base de sondage de l'information sur l'inadmissibilité provenant des enquêtes	187
WALTER MUDRYK et HANSHENG XIE	
Application du contrôle de la qualité à la saisie des données par RIC : Recensement de l'agriculture du Canada de 2001	197
INHO PARK et HYUNSHIK LEE	
Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe	205
JEAN-FRANÇOIS BEAUMONT et ASMA ALAVI	
Estimation robuste par la régression généralisée	217
HUI ZHENG et RODERICK J.A. LITTLE	
Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés	233
FAMING LIANG et ANTHONY YUNG CHEUNG KUK	
Une étude de l'estimation pour population finie au moyen de réseaux neuronaux bayésiens	245
JEROME J. REITER	
Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation	263
Remerciements	
273	
Erratum	
274	

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président	G. J. Brackstone
Membres	D. A. Binder G. J. C. Hole C. Patrick R. Platek (Ancien président)
Rédacteur en chef	M. P. Singh
Rédacteurs associés	M. P. Singh, <i>Statistique Canada</i>

COMITÉ DE RÉDACTION

Rédacteurs adjoints	J.-F. Beaumont, P. Dick, H. Mantel et W. Yung, <i>Statistique Canada</i>
Rédacteurs associés	D. R. Bellhouse, <i>University of Western Ontario</i> D. A. Binder, <i>Statistique Canada</i> J. M. Brick, <i>Westat, Inc.</i> P. Cantwell, <i>U.S. Bureau of the Census</i> J. L. Eltinge, <i>U.S. Bureau of Labor Statistics</i> W. A. Fuller, <i>Iowa State University</i> J. Gambino, <i>Statistique Canada</i> M. A. Hidroglou, <i>Office for National Statistics</i> G. Kalton, <i>Westat, Inc.</i> P. Kot, <i>National Agricultural Statistics Service</i> J. Kovar, <i>Statistique Canada</i> P. Lahiri, <i>JPSM, University of Maryland</i> G. Nathan, <i>Hebrew University</i> D. Norris, <i>Statistique Canada</i> D. Pfeffermann, <i>Hebrew University</i> J. N. K. Rao, <i>Carleton University</i>
Rédacteurs adjoints	T. J. Rao, <i>Indian Statistical Institute</i> J. Reiter, <i>Duke University</i> L.-P. Rivest, <i>Université Laval</i> N. Schenker, <i>National Center for Health Statistics</i> F. J. Scheuren, <i>National Opinion Research Center</i> C. J. Skinner, <i>University of Southampton</i> E. Stasny, <i>Ohio State University</i> D. Steel, <i>University of Wollongong</i> L. Stokes, <i>Southern Methodist University</i> M. Thompson, <i>University of Waterloo</i> Y. Tillé, <i>Université de Neuchâtel</i> R. Valliant, <i>JPSM, University of Michigan</i> J. Wakseberg, <i>Westat, Inc.</i> K. M. Wolter, <i>Iowa State University</i> A. Zaslavsky, <i>Harvard University</i>

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférentiellement en Word au rédacteur en chef, Dr. M. P. Singh, singhmp@statcan.ca (Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Pré Tunney, Ottawa, Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de *Techniques d'enquête* (n° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 30 \$ CA (15 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale des Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Janvier 2005

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

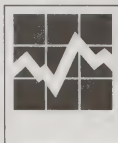
© Ministre de l'Industrie, 2005

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 2004 • VOLUME 30 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



5460





NUMÉRO 2

•

VOLUME 30

•

DÉCEMBRE 2004

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE



